CS794/CO673: Optimization for Data Science

Lec 15: Projection Algorithms

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

November 04, 2022

Problem

Constrained minimization problem:

$$\inf_{\mathbf{w}\in\mathbb{R}^d} \; f(\mathbf{w}) \qquad \text{s.t.} \quad \mathbf{w} \in \bigcap_{i\in I} C_i,$$

- Each $C_i \subseteq \mathbb{R}^d$ is closed, ~~convex~~ and simple
- Projector $\mathrm{P}_i = \mathrm{P}_{C_i}$ can be easily computed
- However, projecting to the intersection C is usually much harder
- Function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex

Problem

Constrained minimization problem:

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

- Each $C_i \subseteq \mathbb{R}^d$ is closed, ~~convex~~ and simple
- Projector $\mathrm{P}_i = \mathrm{P}_{C_i}$ can be easily computed
- However, projecting to the intersection C is usually much harder
- Function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex

Problem

Constrained minimization problem:

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

- Each $C_i \subseteq \mathbb{R}^d$ is closed, ~~convex~~ and simple
- Projector $\mathrm{P}_i = \mathrm{P}_{C_i}$ can be easily computed
- However, projecting to the intersection C is usually much harder
- Function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex

Problem

Constrained minimization problem:

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

- Each $C_i \subseteq \mathbb{R}^d$ is closed, ~~convex~~ and simple
- Projector $\mathrm{P}_i = \mathrm{P}_{C_i}$ can be easily computed
- However, projecting to the intersection C is usually much harder
- Function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex

Problem

Constrained minimization problem:

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

- Each $C_i \subseteq \mathbb{R}^d$ is closed, ~~convex~~ and simple
- Projector $\mathrm{P}_i = \mathrm{P}_{C_i}$ can be easily computed
- However, projecting to the intersection C is usually much harder
- Function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex

Perceptron and SVM revisited

Recall the perceptron problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \; f(\mathbf{w}) \equiv 0$$

$$\text{s.t.} \quad \mathbf{w} \in \bigcap_{i=1}^{n} C_i, \quad \text{where} \quad C_i := \{\mathbf{w} : \langle y_i \mathbf{x}_i, \mathbf{w}\rangle \geq 1\}$$

Similarly, we may rewrite the hard-margin SVM problem as:

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

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

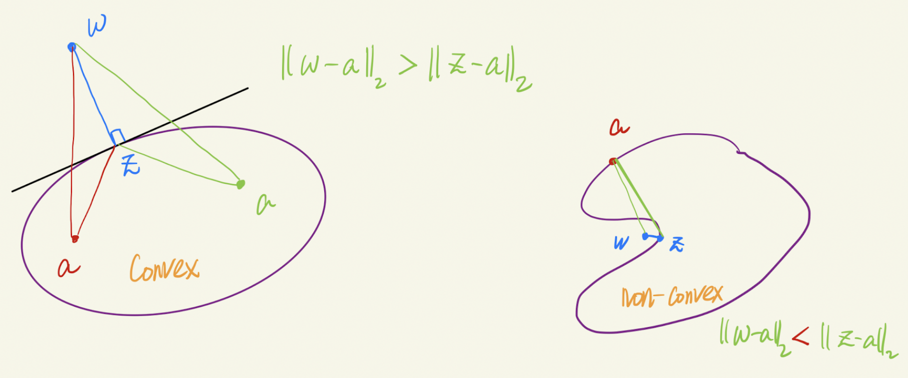
$$\mathrm{P}_{C_i}(\mathbf{z}) := \left[\operatorname*{argmin}_{\mathbf{w}\in C_i} \|\mathbf{w} - \mathbf{z}\|_2\right] = \mathbf{z} + \frac{(1 - \langle y_i \mathbf{x}_i, \mathbf{z}\rangle)_+}{\|\mathbf{x}_i\|_2^2} y_i \mathbf{x}_i.$$

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

Theorem: Fejér's characterization of the closed convex hull

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



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

Algorithmic Significance of Fejér's Result

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 of any problem. 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 there exists some $\mathbf{w}_1$ such that $\|\mathbf{w}_1 - \mathbf{w}\| < \|\mathbf{w}_0 - \mathbf{w}\|$ for all $\mathbf{w} \in C$.

- 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 $\mathrm{dist}(\mathbf{w}_t, C) \to 0$.
- We may also want to prove the convergence (rate) of the whole sequence $\mathbf{w}_t$.

Algorithmic Significance of Fejér's Result

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 of any problem. 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 there exists some $\mathbf{w}_1$ such that $\|\mathbf{w}_1 - \mathbf{w}\| < \|\mathbf{w}_0 - \mathbf{w}\|$ for all $\mathbf{w} \in C$.

- 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 $\mathrm{dist}(\mathbf{w}_t, C) \to 0$.
- We may also want to prove the convergence (rate) of the whole sequence $\mathbf{w}_t$.

Algorithmic Significance of Fejér's Result

Can be used to solve the convex feasibility problem:

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

where the closed (and convex) set $C \subseteq \mathbb{R}^d$ represents the solutions set of any problem. 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 there exists some $\mathbf{w}_1$ such that $\|\mathbf{w}_1 - \mathbf{w}\| < \|\mathbf{w}_0 - \mathbf{w}\|$ for all $\mathbf{w} \in C$.

- 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 $\mathrm{dist}(\mathbf{w}_t, C) \to 0$.
- We may also want to prove the convergence (rate) of the whole sequence $\mathbf{w}_t$.

Algorithmic Significance of Fejér's Result

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 of any problem. 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 there exists some $\mathbf{w}_1$ such that $\|\mathbf{w}_1 - \mathbf{w}\| < \|\mathbf{w}_0 - \mathbf{w}\|$ for all $\mathbf{w} \in C$.

- 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 $\text{dist}(\mathbf{w}_t, C) \to 0$.
- We may also want to prove the convergence (rate) of the whole sequence $\mathbf{w}_t$.

Algorithmic Significance of Fejér's Result

Can be used to solve the convex feasibility problem:

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

where the closed (and convex) set $C \subseteq \mathbb{R}^d$ represents the solutions set of any problem. 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 there exists some $\mathbf{w}_1$ such that $\|\mathbf{w}_1 - \mathbf{w}\| < \|\mathbf{w}_0 - \mathbf{w}\|$ for all $\mathbf{w} \in C$.

- 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 $\mathrm{dist}(\mathbf{w}_t, C) \to 0$.
- We may also want to prove the convergence (rate) of the whole sequence $\mathbf{w}_t$.

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 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:

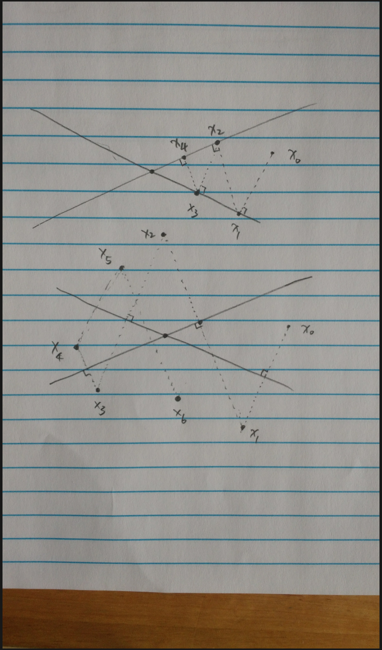
Algorithm 1: Method of alternating projections

Input: $\mathbf{w}_0$

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

  choose set C_{i_t} `// cyclic, random or greedy`

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


x2
x4
x0
x3
x1
x2
x5
x0
x4
x3
x6
x1

Half Justification

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

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

Theorem: Convergence of alternating projections

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

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

L. M. Bregman. "The method of successive projection for finding a common point of convex sets". *Soviet Mathematics Doklady*, vol. 162, no. 3 (1965), pp. 688–692, L. G. Gubin et al. "The Method of Projections for Finding the Common Point of Convex Sets". *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 6 (1967), pp. 1–24. [English translation of paper in *Zh. Vȳchisl. Mat. mat. Fiz.* vol. 7, no. 6, pp. 1211–1228, 1967].

Alternating Bregman Projection

Instead of the Euclidean projection, can also 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.

Algorithm 2: Alternating Bregman projection

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

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

  choose set C_{i_t} `// cyclic, random or greedy`

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

L. M. Bregman. "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 (1966), pp. 1578–1581.

Dyksstra's algorithm

We now present a beautiful algorithm for solving:

$$\min_{\mathbf{w}} \; f(\mathbf{w}) \quad \text{s.t.} \quad \mathbf{w} \in C := \cap_{i \in I} C_i,$$

where f is Legendre and each C_i is closed and convex.

Algorithm 3: Dykstra's algorithm

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

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

- choose set C_{i_t} `// cyclic, random or greedy`
- $\mathbf{w}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{w} \in C_{i_t}} \; f(\mathbf{w}) - \langle \mathbf{w}, \nabla f(\mathbf{w}_t) + \mathbf{a}_{i_t} \rangle$ `// 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`

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

Dykstra = AltMin in the Dual

Apply Fenchel-Rockafellar duality we obtain the dual problem:

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

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

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

- f is Legendre $\implies$ f^* is smooth and convex so AltMin applies

$$\mathbf{w}_{i,t+1}^* = \operatorname*{argmin}_{\mathbf{w}_i^*} f^*\Big(-\mathbf{w}_i^* - \sum_{j\neq i} \mathbf{w}_{j,t}^*\Big) + \sigma_i(\mathbf{w}_i^*)$$

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

Dykstra = AltMin in the Dual

Apply Fenchel-Rockafellar duality we obtain the dual problem:

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

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

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

- f is Legendre $\implies$ f^* is smooth and convex so AltMin applies

$$\mathbf{w}_{i,t+1}^* = \operatorname*{argmin}_{\mathbf{w}_i^*} f^*\Big(-\mathbf{w}_i^* - \sum_{j\neq i} \mathbf{w}_{j,t}^*\Big) + \sigma_i(\mathbf{w}_i^*)$$

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

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:

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

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)

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

where the last inequality follows from (1). Then,

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

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.

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 distribution $\Pi \in \mathbb{R}_+^{m \times n}$ with p and q as marginals such that the transportation cost is minimized:

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

Add a small entropy regularization:

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

W.l.o.g. let $\Pi_0 \propto \exp(-C/\lambda) \geq \mathbf{0}$ and $\mathbf{1}^\top \Pi_0 \mathbf{1} = 1$ to obtain the equivalent problem:

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