LP Theorems

Klejdi Sevdari

February 2022

Affine Spaces

Definition 1 (Affine Space)**.** S is an affine space if $S = L + b$, where L is a linear space and b is an arbitrary vector (offset from the origin).

Definition 2 (Affine Combination of n vectors)**.**

$$y = \sum_{i=1}^{n} \alpha_i x_i,\ \sum_i \alpha_i = 1.$$

Definition 3 (Affine dependence)**.** $x_1, \ldots, x_n$ are affinely dependent if one of them can be written as an affine combination of the others. Or equivalently:

$$\sum_{i=1}^{n} \alpha_i x_i = 0,\ \sum \alpha_i = 0,\ \exists j : \alpha_j \neq 0.$$

Definition 4 (Affine Hull)**.** $\mathsf{aff\,hull}(X) = \cap\{A | A \text{ is an affine space}, X \subseteq A\}$

Theorem 1 (Affine hull = Set of Affine Combinations)**.** $\mathsf{aff\,hull}(X) = \{y | y = \sum_i \alpha_i x_i,\ \sum_i \alpha_i = 1,\ x_i \in X\ \forall i\}$.

Definition 5 (Affine dimension)**.** We define the affine dimension of every set X as the dimension of the affine hull of X. The dimension of an affine space S is the dimension of the largest linear space such that $S = b + L$.

Theorem 2. $\mathsf{dim}(X) = \max k$ *such that* X *contains* $k+1$ *affinely independent vectors.*

Definition 6 (Halfspaces)**.** $\{x \in \mathbb{R}^n | a^T x \leq b\}$ for $a \in \mathbb{R}^n \setminus \{0\}, b \in \mathbb{R}$.

Linear Programming

Definition 7. The standard form is given by:

$$\begin{gathered} \max\ c^T x \\ \text{s.t } Ax \leq b \end{gathered}$$

Definition 8. The equational form is given by:

$$\begin{gathered} \max\ c^T x \\ \text{s.t } Ax = b \\ x \geq 0 \end{gathered}$$

An LP is standard form can be transformed to equational form by:

- $Ax \leq b$ converts to $Ax + z = b$, $z \geq 0$.
- $x \in \mathbb{R}$ converts to $x^+ - x^-,\ x^+ \geq 0, x^- \geq 0$.

Content involving Convex Sets and Affine Spaces

In this section by $co(X)$, we denote the convex hull of X.

Theorem 3 (Lying on one halfspace). *If $\forall x \in X : a^T x \leq b$, then any convex combination y of $x_i \in X$ satisfies $a^T y \leq b$.*

Theorem 4. *If X is finite then $co(X)$ is compact. (This follows from continuity).*

Theorem 5. *If $a^T x = b \; \forall x \in X$, then $a^T y = b \; \forall y \in \mathsf{aff\ hull}(X)$.*

Theorem 6 (Caratheodory). *Let $X \subseteq \mathbb{R}^n$, $\mathsf{dim}(X) = d$. Then $co(X) = \{\sum_{i=1}^{d+1} \alpha_i x_i | x_i \in X, \alpha_i \geq 0, \sum \alpha_i = 1\}$.*

Theorem 7 (Separation thoerem). *Let $C, D \subseteq \mathbb{R}^n$ which are both nonempty, closed, convex, disjoint and C is bounded. Then there is a hyperplane $\{x | a^T x = b\}$ such that:*

$$C \subseteq \{x | a^T x < b\}, D \subseteq \{x | a^T x > b\}.$$

Definition 9 (Polyhedron and Polytope). A polyhedron is the intersection of finitely many halfspaces. A polytope is a bounded polyhedron.

Definition 10 (Cross polytope). $\{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}$.

Definition 11 (Cone). $\mathsf{cone}(X) = \left\{\sum_{i=1}^k \alpha_i x_i : k \in \mathbb{N}_+, x_i \in X, \alpha_i \geq 0\right\}$.

Theorem 8. *Every finitely generated cone is closed.*

Theorem 9 (Farkass Lemma). *$A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$. Exactly one of the following statements is true:*

- $\exists x \in \mathbb{R}^n : Ax = b, \; x \geq 0$,
- $\exists y \in \mathbb{R}^m : y^T A \geq 0^T$ *and* $y^T b < 0$.

An equivalent formulation is the following:
$a_1, \dots, a_n, b \in \mathbb{R}^m$. Exactly one of the following holds:

- $b \in \mathsf{cone}(\{a_1, \dots, a_n\}) = C$,
- *There is a hyperplane separating b from C: $h = \{x \in \mathbb{R}^m : y^T x = 0\}$, $y^T b < 0$ and $y^T a_i \geq 0 \; \forall i$.*

Theorem 10 (Minkowski). *$P \subseteq \mathbb{R}^n$ is a polytope $\iff$ there is a finite set $V \subseteq \mathbb{R}^n$ such that $P = \mathsf{co}(V)$.*

Structure of Polyhedra

Definition 12 (Supporting hyperplane). Let $P \subseteq \mathbb{R}$ be a polyhedron, $c \in \mathbb{R}^n, t \in \mathbb{R}$. Then $h = \{x \in \mathbb{R}^n : c^T x = t\}$ is called a supporting hyperplane if $h \cap P \neq \emptyset$ and $c^T x \leq t, \; \forall x \in P$.

Definition 13 (Face). If P is a polyhedron, and h is a supporting hyperplane, then $F = P \cap h$ is called a face of P. We also call $P, \emptyset$ faces of P. Faces that are not $P, \emptyset$ are called proper faces. Other names include:

- vertex: face of dimension 0,
- edge: face of dimension 1,
- facet: face of dimension $dim(P) - 1$.

Theorem 11 (Vertices and Extreme Points of a Polytope). *Let $P \subseteq \mathbb{R}^n$ be a polytope. Let V denote the set of vertices of P. Let $V_{ext} = \{x \in P : x \notin conv(P \setminus \{x\})\}$ (set of extreme points). Then $V = V_{ext}$ and $P = conv(V)$.*

Theorem 12. *If P is a polyhedron, $V = V_{ext}$.*

Theorem 13. *An intersection of faces is a face.*

Remark: A face is a polyhedron.

Theorem 14. *Let F be a face of P. Then $E \subseteq F$ is a face of P iff E is a face of F.*

Facts about faces of polytopes:

- For any two faces F, G there is a face $M = F \cap G$.
- For any two faces F, G there is a face J such that J contains both F, G and is contained in all faces containing both F and G.
- Every face is the convex hull of its vertices, because each face is a polytope.
- Every face is an intersection of all faces containing it.
- Each face of dimension $dim(P) - 2$ is an intersection of exactly 2 faces.

Definition 14 (Minimal face)**.** A minimal face is a face that does not contain any proper face. All minimal faces are affine spaces.

Facts about faces of polyhydra:

- A polyhedron may not have vertices.
- All minimal faces have the same dimension.
- Every facet is a convex hull of its minimal faces.
- Every minimal face is the intersection of the facets containing it.

Definition 15 (Minimal description of faces)**.** $P = \{x \in \mathbb{R}^n : A'x = b' \text{ and } A''x \leq b''\}$ is described minimally if the omission of any constraint results in changing P and no inequality can be converted to an equality without without changing P.

Theorem 15. *If P is described minimally and $P \neq \emptyset$, $\exists z \in P$ such that $A''z < b$.*

Theorem 16 (Dimension of polyhedra)**.** *If there is $z \in P$ such that $A''z < b''$, then:*

- $dim(P) = n - rank(A')$,
- *z does not belong to a proper face of the polyhedron.*

Theorem 17 (Facets and faces)**.** *Every face of P is a face of some of its facets. There is a one-to-one correspondence between facets of P and inequalities in a minimal descripition of P.*

Corollary 17.1. *Here we list corollaries of Theorem 17:*

- *Every proper face is an intersection of some of its facets.*
- *If $dimP = n$, then its min description is unique up to multiplication of the inequalities and there are no equalities in the minimal description. ($P \subseteq \mathbb{R}^n$).*

Theorem 18 (Description of minimal faces)**.** *F is a minimal face of P $\iff$ $F = \{x : A'x = b', \tilde{A}x = \tilde{b}\}$, where $\tilde{A}x = \tilde{b}$ is a subsystem of $A''x = b''$.*

Theorem 19 (Dimension of Minimal faces)**.** *All minimal faces of P have dimension $n - rank \begin{bmatrix} A' \\ A'' \end{bmatrix}$*

Theorem 20. *The set of optimal solutions (if it exists) is a face. There is always a minimal face E such that $\forall x \in E$, x is an optimal solution.*

Theorem 21. *If P has a vertex, there is always an optimal solution that is a vertex.*

Simplex

LP in equational form:
$\max c^T x$
$Ax = b$
$x \geq 0$
We assume that $Ax = b$ has a solution and that the rows of A are linearly independent.

Definition 16 (Basic feasible solution)**.** x is a basic feasible solution if there exists an m-element set $B \subseteq \{1, \ldots, n\}$ such that:

- A_B is nonsingular.
- $x_j = 0, \ \forall j \notin B$.

Theorem 22. *For each feasible basis, there is a unique basic feasible solution x and x is a vertex. Each vertex corresponds to some basis, sometimes several ones. If the optimum exists, there is also a basic optimal solution.*

Theorem 23. *For each feasible basis, there is exactly one simplex tableau:*

$$Q = -A_B A_N, \ p = A_B^{-1} b, \ z_0 = c_B^T A_B^{-1} b, \ r = c_N - (c_B^T A_B^{-1} A_N)^T$$

Theorem 24. *If $r \leq 0$, the corresponding basic feasible solution is optimal.*

Theorem 25. *The Simplex Method at each step goes from one feasible basis B to another feasible basis B'.*

Duality

Primal LP: max $c^T x$, subject to $Ax \leq b$, $x \geq 0$.
Dual LP: $\min b^T y$, subject to $A^T y \geq c$, $y \geq 0$.

Theorem 26. *The dual of the dual is the primal.*

Theorem 27. *For each feasible x, y to the primal and the dual, it holds: $c^T x \leq b^T y$.*

Corollary 27.1. *If (P) is unbounded, then (D) is infeasible.*

Theorem 28 (Complementary Slackness)**.** *x, y are optimal if $y^T(Ax - b) = 0$ and $(A^T y - c)^T x = 0$.*

Theorem 29. *Let x_0 be such that $Ax_0 \leq b$. Then the following statements are equivalent:*

- *$\forall x$ such that $Ax \leq b$, it holds $c^T x \leq \delta$.*
- *$\exists y \geq 0 : A^T y = c$ and $b^T y \leq \delta$.*

Theorem 30 (Strong Duality)**.** *Exactly one of the following can occur:*

- *Both (P) and (D) are infeasible.*
- *(P) is unbounded and (D) is infeasible.*
- *(P) is infeasible and (D) is unbounded.*
- *(P) and (D) obtain optimal solutions and $c^T x^* = b^T y*$.*

Dualization Recipe

	Primal linear program	Dual linear program
Variables	$x_1, x_2, \ldots, x_n$	$y_1, y_2, \ldots, y_m$
Matrix	A	A^T
Right-hand side	$\mathbf{b}$	$\mathbf{c}$
Objective function	$\max \mathbf{c}^T\mathbf{x}$	$\min \mathbf{b}^T\mathbf{y}$
Constraints	ith constraint has $\leq$ $\geq$ $=$	$y_i \geq 0$ $y_i \leq 0$ $y_i \in \mathbb{R}$
	$x_j \geq 0$ $x_j \leq 0$ $x_j \in \mathbb{R}$	jth constraint has $\geq$ $\leq$ $=$

Figure 1: Dualization Recipe

ILP

$\max c^T x = b$
$Ax \leq b$
x is integer

Theorem 31. *Week duality holds.* $\delta = \sup\{c^T x \mid Ax \leq b,\ x \text{ is integer}\}$,
$\gamma = \inf\{b^T y \mid y^T A = c, y \geq 0,\ y \text{ is integral}\}$

Definition 17. ILP of Knapsack:
$\max c^T x$
s.t. $a^t x \leq b$
$x \in \{0, 1\}$

Definition 18. LP relaxation:
$\max c^T x$
s.t. $a^t x \leq b$
$x \leq 1$
$x \geq 0$

Theorem 32. *Let x^* be an optimal solution of the relaxation, which is a vertex of the feasible region. Then, there is at most one coordinate $x_i^* \notin \{0, 1\}$.*

Theorem 33. *Let $x*$ be an optimal solution of the LP relaxation. Then define two integral solutions:*

$$\overline{x_i} = \begin{cases} 1 & x_i^* = 1 \\ 0 & \textit{otherwise} \end{cases} \quad \hat{x}_i = \begin{cases} 1 & x_i^* \textit{ is fractional} \\ 0 & \textit{otherwise} \end{cases}$$

Then $\overline{x}$ and $\hat{x}$ are feasible to ILP and of them achieves at least half of the optimal ILP objective.

Definition 19 (Total Unimodularity)**.** A is TU if each square submatrix of A has determinant $\{0, 1, -1\}$.

Theorem 34. *If A is totally unimodular, then so is A^T, $\begin{bmatrix} A \\ I \end{bmatrix}$, $\begin{bmatrix} A^T \\ -I \end{bmatrix}$.*

Theorem 35. *Let A be TU and b any vector in $\mathbb{Z}^m$. Then each vertex of $P = \{x : Ax \leq b\}$ is an integer vector.*

Theorem 36 (Cramer's rule)**.** *If $Ax = b$, then $x_i = \det(A_{i \to e_i}) / \det(A)$.*

Definition 20 (Integer Polyhedron)**.** P is an integer polyhedron if for each vector c, such that $\max\{c^T x | x \in P\}$ is finite, the maximum is attained by some integer vector.

Corollary 36.1. *P is an integer polyhedron if all vertices are integer.*

Corollary 36.2. *Let A be TU and $b \in \mathbb{Z}^m$. Then $P = \{x | Ax \leq b\}$ is an integer polyhedron.*

Corollary 36.3. *Let A be TU and $b \in \mathbb{Z}^m$ and $c \in \mathbb{Z}^n$. If (P) and (D) have finite optima, then they have integer optimum solutions.*

Definition 21 (Perfect Matching)**.** A matching that covers all vertices of the graph.

Definition 22 (Neighborhood)**.** $G = (V, E),\ A \subseteq V,\ N(A) = \{w \in V \setminus A : vw \in E \text{ for some } v \in A\}$.

Definition 23 (Hall's Theorem)**.** Let $G = (V, E)$ be a bipartite graph with bipartition $V = X \cup Y$. G has a matching which covers X iff $|N(T)| \geq |T|$, $\forall T \subseteq X$.

Definition 24. For $F \subseteq E : \chi_e^F = \begin{cases} 1 & e \in F \\ 0 & \text{otherwise} \end{cases}$.

Definition 25 (Matching Polytope)**.** $P_{\text{matching}}(G) = co\{\chi^M | M \text{ matching in } G\}$

Theorem 37. *If G is bipartite and A its incidence matrix, then*

$$P_{matching}(G) = \{x \in \mathbb{R}^E | Ax \leq 1, x \geq 0\}$$

Definition 26 (Independent set polytope)**.**

$$P_{\text{ind, set}}(G) = co\{\chi^I | I \subseteq V \text{ independent set in } G\} = \{y \in \mathbb{R}^V | A^T y \leq 1, y \geq 0\} \text{ if G bipartite.}$$

BP Matching

Theorem 38. *Let G be a bipartite graph. Then the max cardinality of a matching in G is equal to the minimum cardinality of a vertex cover in G.*

Note: Incidence Matrix of a Bipartite Graph is totally unimodular. Thus, problems like max independent set, maximum matching can be solved by their LP relaxation efficiently.

Definition 27 (Neighborhood)**.** $G = (V, E), A \subseteq V$. $N(A) = \{w \in V \setminus A : (v, w) \in E \text{ for some } v \in A\}$.

Theorem 39 (Hull's Theorem)**.** *Let $G = (V, E)$ be a bipartite graph with bipartition $V = X \cup Y$. G has a matching that covers X iff $|N(T)| \geq |T|\ \forall T \subseteq X$.*

Directed Graphs and Flows

Definition 28 (Incidence Matrix of a directed graph)**.** $M \in \mathbb{R}^{V \times E}$. $M_{v,e} = 1$ if e enters v. $M_{v,e} = -1$ if e leaves v and it is 0 otherwise. M is TU.

Definition 29 (Circulation and Flow)**.** $f \in \mathbb{R}^E$ is called a circulation if $Mf = 0$. f is a flow if $M'f = 0$ where M' is M with rows of vertices s and t removed.

Definition 30 (Max Flow in a directed Graph)**.** $\max w^T x$ subject to $M'x = 0, x \leq c, x \geq 0$.

Note: I stress again that M is TU. Thus the solutions will be integral vectors, if c is integral. We can also add lower bounds ($x \geq d$) to the Max flow primal with d integral, and the solutions remain integral.

Theorem 40. *If $\alpha = \max\{w^T x | x \in P\}$ for any w and integer polyhderon P, then $P' = \{x | x \in P, w^T x = \alpha\}$ is a face of P and thus also an integer polyhedron.*

Definition 31 (Min Cost Flow)**.** Given directed graph $G = (V, E)$, $s, t \in V$, $c, \kappa \in \mathbb{R}^E_+$ a maximum $s - t$ flow is called min cost flow if it minimizes $\kappa^T x$ over all $s - t$ flows of maximum size.

Definition 32 (Integrality Gap)**.** Let x^I be the optimal solution of the ILP and x^* the optimal solution of its relaxation. We call integrality gap (for LP maximization problems) the ratio $\frac{c^T x^*}{c^T x^I}$.

Theorem 41. *If G is not bipartite, the integrality gap of a vertex cover is at most 2.*

Max Cardinality Bipartite Matching

Definition 33 (Augmenting Path)**.** Let M be a matching. Path $= (v_0, v_1, \ldots, v_t)$ is M-augmenting if:

- t is odd,
- $(v_1, v_2), (v_3, v_4), \ldots, (v_{t-2}, v_{t-1}) \in M$,
- v_0, v_t not matched in M.

Theorem 42. *Let G be a graph (not necessarily bipartite) and M a matching in G. Then either M is a max-cardinality matching or there is an M-augmenting path.*

Algorithm: Finding Augmenting Path in Bipartite Graphs.

```
# input: G=(V, E) with bipartition L, R. matching M
for e in E:
    u,v = e # u in L, v in R
    if e in M:
        orient edge from v to u # all matched edges go from R to L
    else: orient edge from u to v
L' = vertices of L unmatched, R' = ...
if there is oriented path from L' to R':
    return path
```

Theorem 43. *Max Cardinality Matching can be found in $O(n)$ executions of finding augmenting path.*

Min Cost Perfect Bipartite Matching

Definition 34.

$$\delta(S) = \{e \in E : |e \cap S| = 1\}, \; \forall S \subseteq V$$

Definition 35. (Min Cost Perfect BP Matching LP)

$$\min \sum_{e \in E} x_e$$

$$\text{such that } \; \forall v : \sum_{e \in \delta(v)} x_e = 1 \text{ and } x_e \geq 0$$

Remark: We can safely assume that $c \geq 0$*.*

Definition 36.

$$E_=(y) = \{(uv) \in E : y_u + y_v = c_{uv}\}$$

Definition 37 (Alternating tree)**.** $T = (V(T), E(T), r)$ is a tree with root r and T is a subgraph of G. $A(T)$ are the odd levels and $B(T)$ are the even levels.

Definition 38 (M-Alternating Tree)**.** T is M-alternating if each $u \in A(T)$ has exactly one son v and $(uv) \in M$.

In Appendix A you find the min cost perfect bipartite matching Algorithm.

Theorem 44. *Throughout the aforementioned algorithm the two invariants are maintained:*

- y *is dual feasible.*
- M *and* y *satisfy Complementary Slackness conditions.*

This guarantees the correctness of the algorithm.

Perfect Non-Bipartite Matching

Definition 39 (Shrinking)**.** Let X, Y be sets. We define the shrinking operator in the following way:

$$X/Y = \begin{cases} X & X \cap Y = \emptyset \\ (X \setminus Y) \cup \{\mathcal{Y}\} & \text{otherwise, where } \mathcal{Y} \text{ is a representative single element} \end{cases}$$

In the context of graphs, we define V/C by deleting the vertices in C and creating a new representative vertex $\mathcal{C}$. The edges that had both their endpoints in C are deleted, and the edges with one endpoint to some element in C have now an endpoint in $\mathcal{C}$ (and the other endpoint is not modified).

Theorem 45. *If* $C \subseteq V$ *is an odd cycle and* M' *a perfect matching in* G/C*. Then there exists* $M \supseteq M'$ *perfect matching in* G*.*

Blossom Algorithm:

```
# M = Empty, G' = G
find r unmatched, initialize Tree with root r
find uv in E, u in B(T), v not in T:
   either increase M or T
if there exists uv in R with u in B(T) and v in B(T):
   P_u, P_v = paths to common ancestor
   C = Union(P_u, P_v)
   G' = G/C, T = T/C, M=M/C
```

Matroids, Submodular Functions and Greedy Algorithms

Definition 40 (Matroid)**.** Let X be a finite set, $\mathcal{I} \subseteq 2^X$. $(X, \mathcal{I})$ is a matroid if:

- $\emptyset \in \mathcal{I}$
- $Y \in \mathcal{I}$, $Z \subseteq Y$ implies $Z \in \mathcal{I}$.
- $Y, Z \in \mathcal{I}$, $|Y| < |Z|$ implies $\exists z \in Z$ such that $Y \cup \{z\} \in \mathcal{I}$.

Definition 41 (Basis)**.** B is a basis of $\mathcal{I}$ if $\forall B' \in \mathcal{I}$, $B \subseteq B'$ implies $B = B'$. Any two bases have the same cardinality.

Greedy Algorithm in full Generality
Set up: $(X, \mathcal{I}), w : X \to \mathbb{R}, W(Y) = \sum_y w(y)$.

Greedy ALG:

```
Select y not in {y_1, ..., y_r} such that:
    {y_1, ..., y_r, y} in I
    w(y) is as large as possible
Repeat until no y is available anymore # Reach a basis
```

Theorem 46 (Optimality of Greedy Algorithms)**.** *If $(X, \mathcal{I})$ is a matroid, then the described Greedy Algorithm finds basis of max weight.*

Definition 42 (Submodular Functions)**.** Let X be a finite set and $f : 2^X \to \mathbb{R}$. Then f is submodular if any of the following 2 equivalent conditions hold:

- $\forall A, B \subseteq X, f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$,
- $\forall A \subseteq B \subseteq X, f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$, $\forall x \notin B$.

Remarks:

- Linear functions of the form $f(A) = \sum_{a \in A} f(a)$ are submodular. Its negation $-f$ is also submodular.
- If f, g are submodular, then so is $\alpha f + \beta g$ for $\alpha, \beta \geq 0$.

Lemma 47. *If B_1, B_2 are two bases of a matroid, then there exists a bijection $\alpha : B_1 \to B_2$ such that $(B \setminus b) \cup \{\alpha(b)\}$ is a basis $\forall b \in B_1$.*

Monotone Submodular Maximization
Greedy ALG for Monotone Submodular functions:

```
S = empty, A = {a: Union(S, a) in I}
while A != empty:
    e = argmax{f_s(a): a in A}
    S = Union(S, e)
    A = {e: Union(S, e) in I}
return S
```

Note: $f_S(a) = f(S \cup \{a\}) - f(S)$.

Theorem 48 (Approximation Algorithm for monotone submodular functions)**.** *The previous algorithm gives a $\frac{1}{2}$-approximation for maximizing a monotone submodular function under Matroid constraints.*

Total Dual Integrality

Definition 43 (TDI)**.** A system $Ax \leq b$ is TDI if $\min\{b^T y : A^T y = c, y \geq 0\}$ for all integral c such that this optimum exists and is finite.

Theorem 49 (TDI and the Primal)**.** *If $Ax \leq b$ is TDI and b integral, then $\{x : Ax \leq b\}$ is an integral polyhedron.*

Theorem 50. *Any rational polyhedron P can be written as $P = \{x : Ax \leq b\}$ where:*

- *A is integral*
- *$Ax \leq b$ is TDI.*

Moreover, b is integral $\iff$ P is integral.

Warning: Being TDI is a property of the system of inequalities, not of the polyhedron.

Convex Optimization Basics

Theorem 51 (Cauchy Schwarz). *$|u^T v| \leq \|u\|\|v\|$. From this we get $\cos(\alpha) = \frac{u^T v}{\|u\|\|v\|}$.*

Definition 44 (Spectral Norm).
$$\|A\| = \max_{\|v\|=1} \|Av\|$$

Theorem 52 (Lipshitz and Differntiable Functions). *If $X \neq \emptyset$, X is open, f differentiable over X. Then TFAE:*

- *f is B-Lipschitz: $\|f(x) - f(y)\| \leq B\|x - y\|$, $\forall x, y \in X$.*
- *$\|\nabla f(x)\| \leq B$, $\forall x \in X$.*

Theorem 53 (Convexity is almost Continuity). *If f is convex over an open domain, then f is continuous.*

Theorem 54 (1^{st} order characterization of convexity). *If f is differentiable and convex, then:*

- *$f(y) \geq f(x) + \nabla f(x)^T (y - x)$, $\forall x, y \in domf$*
- *$(\nabla f(y) - \nabla f(x))^T (y - x) \geq 0$, $\forall x, y \in domf$.*

Theorem 55 (2^{nd} order characterization of convexity). *If f is twice differentiable, then f is convex iff $\nabla^2 f(x) \succeq 0$ (The Hessian is semi positive definite).*

Theorem 56 (Unconstrained Minimization of Convex Function). *If x is such that $\nabla f(x) = 0$, then x is a global minimizer.*

Theorem 57 (Constrained Minimization). *x^* is a minimum of f over the convex set X iff $\nabla f(x^*)^T (x - x^*) \geq 0$.*

Gradient Descent

Definition 45 (Gradient Descent Step). $x_{t+1} = x_t - \gamma \nabla f(x_t)$.

Definition 46 (Convenient Definition). $g_t = \nabla f(x_t) = \frac{1}{\gamma}(x_t - x_{t+1})$.

Definition 47 (Useful bound). $f(x_t) - f(x^*) \leq \nabla f(x_t)^T (x_t - x^*) = g_t^T (x_t - x^*)$. (follows by convexity)

Theorem 58 (Cosine Theorem and Vanilla Analysis). *$2v^T w = \|v\|^2 + \|w\|^2 - \|v - w\|^2$. This gives:*

$$g_t^T (x_t - x^*) = \frac{\gamma}{2}\|g_t\|^2 + \frac{1}{2\gamma}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$$

.
Summing up, we get:

$$\sum_{t=0}^{T-1} g_t^T (x_t - x^*) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1} \|g_t\|^2 + \frac{1}{2\gamma}\left(\|x_0 - x^*\|^2 + \|x_T - x^*\|^2\right)$$

.
Dropping the negative term above and using the "useful bound", we get the following bound for the average error:

$$\textit{Average error} = \frac{1}{T}\sum_{t=0}^{T-1} (f(x_t) - f(x^*)) \leq \frac{\gamma}{2T}\sum_{t=0}^{T-1} \|g_t\|^2 + \frac{1}{2T\gamma}\|x_0 - x^*\|^2$$

Theorem 59 (Lipschitz GD). *Suppose f is convex, differentiable, B-Lipschitz ($\nabla f(x) \leq B$). Choosing:*

$$\gamma = \frac{R}{B\sqrt{T}} \textit{ gives: Average Error } \leq \frac{RB}{\sqrt{T}}$$

Definition 48 (Smooth Convex Functions)**.** Let f be differentiable and convex. f is L-smooth over X if:

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|x-y\|^2$$

The next 3 theorems are not very important for the course material.

Theorem 60 (Function L-smooth implies Gradient L-Lipschitz)**.** *f is L-smooth iff $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$.*

Theorem 61. *$f_1, \ldots, f_m$. f_i is L_i-smooth. Then $f = \sum_i \lambda_i f_i$ is $(\sum_i \lambda_i L_i)$-smooth, where all $\lambda_i \geq 0$.*

Theorem 62. *Let $f(x)$ be L-smooth. Then $f(Ax+b)$ is $L\|A\|^2$-smooth.*

Lemma 63 (GD steps always improve on L-smooth functions)**.** *Let f be L-smooth and $\gamma = \frac{1}{L}$. Then GD satisfies:*

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$$

Theorem 64. *$f : \mathbb{R}^d \to \mathbb{R}$ convex, differentiable with global minimum x^*. Moreover f is L-smooth with stepsize $\gamma = \frac{1}{L}$, GD achieves*

$$f(x_T) - f(x^*) \leq \frac{L}{2T}\|x_0 - x^*\|^2$$

Definition 49 (Strong Convexity)**.** f convex, differentiable. $\eta > 0$. f is called η-strongly convex over X if

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\eta}{2}\|x-y\|^2$$

.

Theorem 65. *Let f be convex, differentiable, L-smooth, η-strongly convex and let x^* be its unique global minimizer. Choosing $\gamma = \frac{1}{L}$, GD satisfies:*

- $\|x_{t+1} - x^*\|^2 \leq (1 - \frac{\eta}{L})\|x_t - x^*\|^2$,
- $f(x_T) - f(x^*) \leq \frac{L}{2}(1 - \frac{\eta}{L})^T\|x_0 - x^*\|^2$. *Note that the T in the left hand side is a power, not a transpose.*

Projected Gradient Descent

Definition 50 (Projected GD step)**.** $y_{t+1} = x_t - \gamma g_t \to x_{t+1} = \prod_X(y_{t+1})$ = Convex Projection of y_{t+1}.

Theorem 66. *$d_y(x) = \|x-y\|^2$ is strongly convex and achieves a unique maximum over closed, convex set X. (The Convex projection is well defined.)*

Lemma 67. *Let X closed, convex. $x \in X, y \in \mathbb{R}^d$ Let $y^* = \prod_X(y)$. TFAE:*

- $(x - y^*)^T(y - y^*) \leq 0$,
- $\|x - y^*\|^2 + \|y - y^*\|^2 \leq \|x - y\|^2$.

Theorem 68 (Lipschitz Projected GD)**.** *Let f be convex, differentiable. Let X be closed and convex. x^* a minimizer of f over X. Suppose $\|x_0 - x^*\| \leq R$ and $\|\nabla f(x)\| \leq B$, $\forall x \in X$. Choosing stepsize $\gamma = \frac{R}{B\sqrt{T}}$ projected GD gets:*

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{RB}{\sqrt{T}}$$

Lemma 69 (Smooth Functions)**.** *f differentiable and smooth with parameter L over a closed and convex set X. With $\gamma = \frac{1}{L}$, projected GD with arbitrary $x_0 \in X$ satisfies:*

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|f(x_t)\|^2 + \frac{L}{2}\|y_{t+1} - x_{t+1}\|^2$$

Theorem 70. *f convex, differentiable, X closed, convex. Assume there is a minimizer x^* of f over X and that f is $L-smooth$. With $\gamma = \frac{1}{L}$, projected GD satisfies:*

$$\textit{Average Error} = f(x_T) - f(x^*) \leq \frac{L}{2T}\|x_0 - x^*\|^2$$

Subgradient Descent

Definition 51 (Subgradient)**.** $g \in \mathbb{R}^d$ is called subgradient of f at x if

$$f(y) \geq f(x) + g^T(y-x),\ \forall y \in \mathrm{dom} f$$

Theorem 71. *f convex, $domf$ open. Then $\|g\| \leq B$, $\forall x \in domf$ iff f is B-Lipschitz.*

Definition 52 (Subgradient Step)**.** Choose $g_t \in \partial f(x_t)$. $x_{t+1} = x_t - \gamma_t g_t$. (Gamma is usually constant)

Theorem 72 (Lipschitz Subgradient Descent)**.** *f convex and B-Lipschitz with global minimum x^*. Assume $\|x_0 - x^*\| \leq R$. With $\gamma = \frac{R}{B\sqrt{T}}$, subGD achieves:*

$$\textit{Average Error} \leq \frac{RB}{\sqrt{T}}$$

Warning: Smoothness does not make sense for non differentiable functions.

Definition 53 (Strong Convexity: Non differentiable case)**.** f convex, $\eta > 0$. f is called η-strongly convex if

$$f(y) \geq f(x) + g^T(y-x) + \frac{\eta}{2}\|y-x\|^2,\ \forall x, y \in \mathrm{dom} f,\ \forall g \in \partial f(x)$$

Theorem 73 (SubGD with Strong Convexity)**.** *f η-strongly convex, x^* unique minimum of f. With $\gamma_t = \frac{2}{\eta(t+1)}$, SubGD achieves*

$$f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t x_t\right) - f(x^*) \leq \frac{2B^2}{\eta(T+1)} \qquad \textit{where } B = \max_t \|g_t\|$$

Stochastic Descent

The function to optimize is $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$.

Definition 54 (Stochastic Gradient Descent Step)**.** Sample $i \in \{1, \ldots, n\}$ at random. $x_{t+1} = x_t - \alpha_t \nabla f_i(x_t)$.

Definition 55 (Convenient Definition)**.** $g_t = \nabla f_i(x_t)$ (Notice the i subindex).

Theorem 74. *g_t is an unbiased estimator of $\nabla f(x_t) = \nabla \sum_i f_i(x_t)$.*

Note: In the proof of this theorem it is shows that $f(x_t) - f(x^*) \leq g_t^T(x_t - x^*)$ holds in expectation.

Theorem 75. *f convex, differentiable, x^* global minimizer, $\|x_0 - x^*\| \leq R$, $\mathbb{E}[\|g_t\|^2] \leq B^2$ $\forall t$. With $\gamma = \frac{RB}{\sqrt{T}}$, SGD achieves:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[f(x_t)\right] - f(x^*) \leq \frac{RB}{\sqrt{T}}$$

Theorem 76. *f differentiable, η-stongly convex. With $\gamma_t = \frac{2}{\eta(t+1)}$, Stochastic GD satisfies:*

$$\mathbb{E}\left[f\left(\frac{2}{T(T+1)}\sum t x_t\right)\right] - f(x^*) \leq \frac{2B^2}{\eta(T+1)}$$

Online Optimization

The setting: $f(x) = \sum_{t=1}^{T} f_t(x)$, f_t convex, X convex.
At time t, choose x_t incur loss $f_t(x_t)$.
Goal: Minimize Regret $R_T = \sum_{t=1}^{T} f_t(x_t) - f(x^*)$ where x^* is a static optimum.

Reducing Online Convex Optimization to Online Linear Optimization:
At each step, we receive loss vector $l_t = \nabla f_t(x_t)$ and incur loss $l_t^T x_t$.

Definition 56 (Online GD step)**.** $y_{t+1} = x_t - \gamma l_t$, $x_{t+1} = \prod_X(y_{t+1})$.

Theorem 77. *Let $R = \max\|x - x^*\|$ and assume that $\|l_t\| \leq B$, $\forall t$. With $\gamma = \frac{R}{B\sqrt{T}}$, OGD achieves regret $R_T \leq RB\sqrt{T}$.*

Learning from Expert Advice

The setting: n experts, In timestep $t \in \{1, \ldots, T\}$:

- choose $i \in \{1, \ldots, n\}$ and follow advice of expert i,
- Observe loss vector $l_t \in [-1, 1]^n$, incur loss $l_{t,i}$.

Modelling the Problem:
Our choices are $X = \{e_1, \ldots, e_n\} \subseteq \mathbb{R}^n$. X is not convex, so we have to consider a probability simplex Δ defined as:

$$\Delta = \{x : x \geq 0, \|x\|_1 = 1\}$$

Then define the expected regret as:

$$\mathbb{E}[R_T] = \sum_{t=1}^{T} l_t^T x_t - \sum_{t=1}^{T} l_t(i^*) \leq \sum l_t^T x_t - \sum l_t^T x^* \leq RB\sqrt{T}$$

where $x^* = \arg\min_{x \in \Delta} \sum l_t^T x$, $R = \max_{i,j} \|e_i - e_j\|^2 = \sqrt{2}$, $B = \max_t \|l_t\|^2 \leq \sqrt{n}$.

Follow the Regularized Leader

We first choose a convex function Φ which is independent on the input. Then we choose $x_0 = \arg\min_{x \in X} \Phi(x)$. Then at time t:

$$x_t = \arg\min_{x \in X} \sum_{j=1}^{t-1} l_j^T x + \gamma \Phi(x)$$

Definition 57 (Dual Norm)**.** Let $\| \circ \|$ be a norm. Its dual norm is defined as $\|x\|_* = \max\{z^T x : \|z\| \leq 1\}$.

Remark: The dual of $\| \circ \|_2$ is $\| \circ \|_2$. The dual of $\| \circ \|_1$ is $\| \circ \|_\infty$.

Theorem 78. *If Φ is 1-strongly convex with respect to $\| \circ \|$ on X (meaning $\Phi(y) \geq \Phi(x) + \nabla\Phi(x)^T(y - x) + \frac{1}{2}\|x - y\|^2$), then FTRL achieves regret*

$$R_T \leq 2\gamma \sum \|l_t\|_*^2 + \frac{1}{\gamma}\left(\Phi(x^*) - \Phi(x_0)\right)$$

Corollary 78.1 (l_2 Regularization)**.** *$\Phi(x) = \frac{1}{2}\|x\|_2^2$ is 1-strongly convex with respect to $\|\|_2$. In the expert setting with $X = \Delta$, $x_0 = (\frac{1}{n}, \ldots, \frac{1}{n})$, we have that $\|x\|_2^2 \leq 1 \forall x \in \Delta$ which gives $\Phi(x^*) - \Phi(x_0) \leq 1$. We also have $l_t \in [-1, 1]^n$ which gives $\|l_t\| \leq \sqrt{n}$. With $\gamma = \frac{1}{\sqrt{nT}}$, we achieve*

$$R_t \leq 2\sqrt{2Tn}$$

Entropy Regularization of FTRL

Definition 58 (Negative Entropy)**.** $\Phi(x) = \sum x_i \log(x_i)$. Φ is 1-strongly convex with respect to the $\| \circ \|_1$ norm over the probability simplex in $\mathbb{R}^n$. Some properties:

$$\nabla\Phi(x) = \mathbb{1} + \begin{pmatrix} \log x_1 \\ \vdots \\ \log x_n \end{pmatrix} \qquad \arg\min_{x \in \Delta_n} \Phi(x) = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

Theorem 79. *FTRL over X with entropy regularization achieves regret $R_T \leq 2RB\sqrt{2T}$ where $R^2 = \max_{x \in X} \Phi(x) - \Phi(x_0)$ and $B = \max_t \|l_t\|_\infty$.*

Corollary 79.1. *$X = \Delta \subseteq \mathbb{R}^n$, $B = 1$ because $l_t \in [-1, 1]^n$, $R^2 = \log n$. We get regret $2\sqrt{2T \log n}$.*

Definition 59 (Convenient Definition)**.** $H(x) = \gamma \left(\sum_{t=1}^{t} l_i\right)^T x + \Phi(x)$.
We want $x_{t+1} = \arg\min_{x\in\Delta} H(x)$. One condition that is sufficient for this to hold is for $\nabla H(x)$ to be perpendicular to Δ.

Hedge ALG:
$w_1 = \mathbb{1}$,
$w_{t+1}(j) = w_t(j)\exp(-\gamma l_t(j)),\ \forall j = 1, \ldots, n$
$x_{t+1} = \frac{w_{t+1}}{\|w_{t+1}\|_1}$.

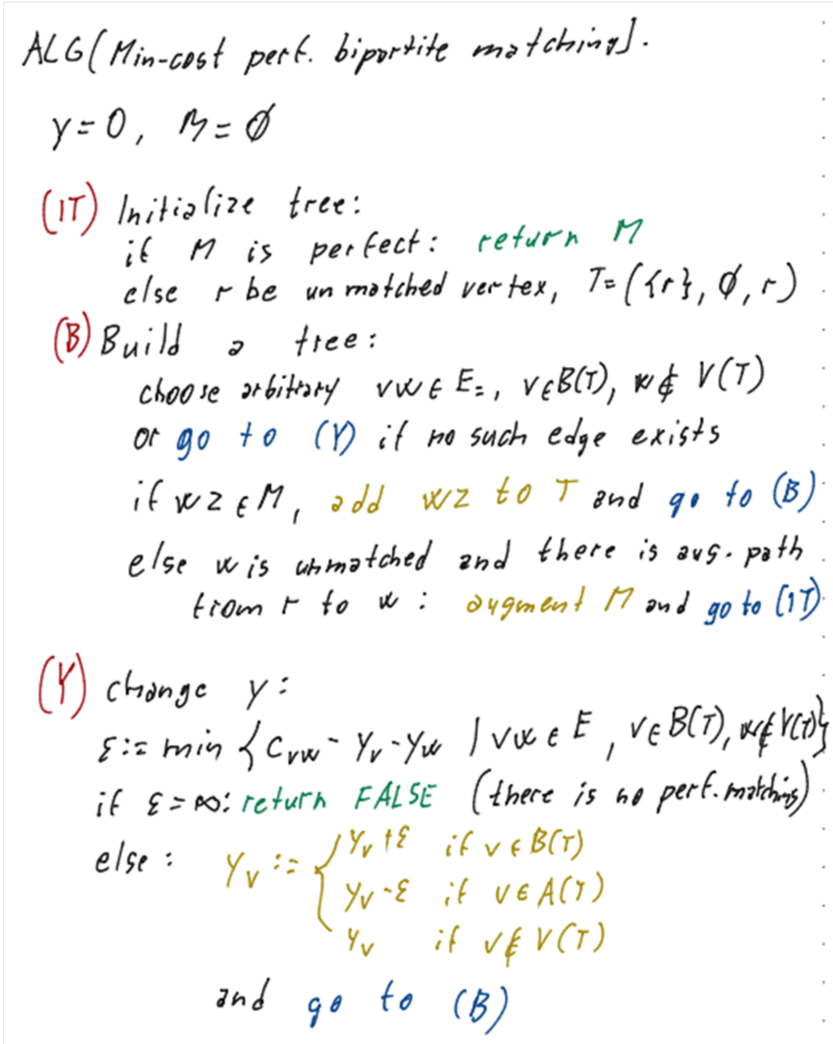
Theorem 80. *Hedge ALG achieves regret*

$$R_T = \sum l_t^T x_t - \sum l_t e_{i^*} \leq \gamma \sum \|l_t\|_\infty^2 + \frac{1}{\gamma}\log n$$

if $\|l_t\|_\infty \leq B$ *we can choose:*

$$\gamma = \frac{\sqrt{\log n}}{B\sqrt{T}} \text{ achieving } R_T \leq 2B\sqrt{T\log n}.$$

Appendix A



ALG(Min-cost perf. bipartite matching).

$y = 0$, $M = \emptyset$

(IT) Initialize tree:
if M is perfect: return M
else r be unmatched vertex, $T = (\{r\}, \emptyset, r)$

(B) Build a tree:
choose arbitrary $vw \in E_=$, $v \in B(T)$, $w \notin V(T)$
or go to (Y) if no such edge exists
if $wz \in M$, add wz to T and go to (B)
else w is unmatched and there is aug. path from r to w: augment M and go to (IT)

(Y) change y:
$\varepsilon := \min\{c_{vw} - y_v - y_w \mid vw \in E, v \in B(T), w \notin V(T)\}$
if $\varepsilon = \infty$: return FALSE (there is no perf. matching)
else: $y_v := \begin{cases} y_v + \varepsilon & \text{if } v \in B(T) \\ y_v - \varepsilon & \text{if } v \in A(T) \\ y_v & \text{if } v \notin V(T) \end{cases}$
and go to (B)

Figure 2: Min Cost Perfect Bipartite Matching Algorithm