Submodular Functions, Optimization, and Applications to Machine Learning

— Fall Quarter, Lecture 2 —

http://www.ee.washington.edu/people/faculty/bilmes/classes/ee563_spring_2018/

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
http://melodi.ee.washington.edu/~bilmes

Oct 5th, 2020

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$$

Cumulative Outstanding Reading

- Read chapter 1 from Fujishige's book.

Class Road Map - EE563

- L1(3/26): Motivation, Applications, & Basic Definitions,
- L2(3/28): Machine Learning Apps (diversity, complexity, parameter, learning target, surrogate).
- L3(4/2): Info theory exs, more apps, definitions, graph/combinatorial examples
- L4(4/4): Graph and Combinatorial Examples, Matrix Rank, Examples and Properties, visualizations
- L5(4/9): More Examples/Properties/ Other Submodular Defs., Independence,
- L6(4/11): Matroids, Matroid Examples, Matroid Rank, Partition/Laminar Matroids
- L7(4/16): Laminar Matroids, System of Distinct Reps, Transversals, Transversal Matroid, Matroid Representation, Dual Matroids
- L8(4/18): Dual Matroids, Other Matroid Properties, Combinatorial Geometries, Matroids and Greedy.
- L9(4/23): Polyhedra, Matroid Polytopes, Matroids $\rightarrow$ Polymatroids
- L10(4/29): Matroids $\rightarrow$ Polymatroids, Polymatroids, Polymatroids and Greedy,
- L11(4/30): Polymatroids, Polymatroids and Greedy
- L12(5/2): Polymatroids and Greedy, Extreme Points, Cardinality Constrained Maximization
- L13(5/7): Constrained Submodular Maximization
- L14(5/9): Submodular Max w. Other Constraints, Cont. Extensions, Lovasz Extension
- L15(5/14): Cont. Extensions, Lovasz Extension, Choquet Integration, Properties
- L16(5/16): More Lovasz extension, Choquet, defs/props, examples, multiliear extension
- L17(5/21): Finish L.E., Multilinear Extension, Submodular Max/polyhedral approaches, Most Violated inequality, Still More on Matroids, Closure/Sat
- L–(5/28): Memorial Day (holiday)
- L18(5/30): Closure/Sat, Fund. Circuit/Dep
- L19(6/6): Fund. Circuit/Dep, Min-Norm Point Definitions, Review & Support for Min-Norm, Proof that min-norm gives optimal, Computing Min-Norm Vector for B_f maximization.

Last day of instruction, June 1st. Finals Week: June 2-8, 2018.

Class Road Map - EE563

- L1(9/30): Motivation, Applications, Definitions, Properties
- L2(10/5): Sums concave(modular), uses (diversity/costs, feature selection), information theory, Monge matrices, graph examples.
- L3(10/7):
- L4(10/12):
- L5(10/14):
- L6(10/19):
- L7(10/21):
- L8(10/26):
- L9(10/28):
- L10(11/2):
- L11(11/4):
- L12(11/9):
- L–(11/11): Veterans Day, Holiday
- L13(11/16):
- L14(11/18):
- L15(11/23):
- L16(11/25):
- L17(11/30):
- L18(12/2):
- L19(12/7):
- L20(12/9): maximization.

Last day of instruction, Fri. Dec 11th. Finals Week: Dec 12-18, 2020

Two Equivalent Submodular Definitions

Definition 2.2.1 (submodular concave)

A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \tag{2.7}$$

An alternate and (as we will soon see) equivalent definition is:

Definition 2.2.2 (diminishing returns)

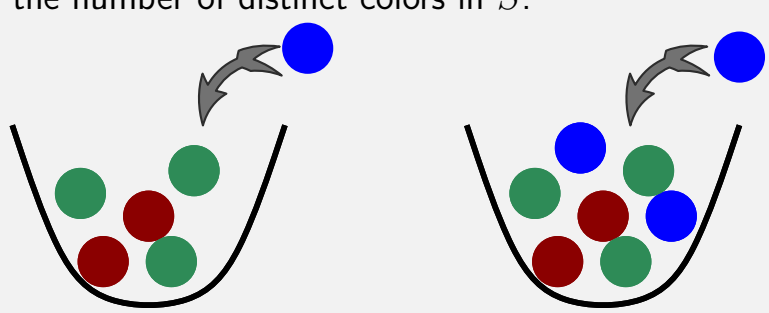
A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A \subseteq B \subset V$, and $v \in V \setminus B$, we have that:

$$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B) \tag{2.8}$$

• The incremental "value", "gain", or "cost" of v decreases (diminishes) as the context in which v is considered grows from A to B.
• Gain notation: Define $f(v|A) \triangleq f(A + v) - f(A)$. Then function f is submodular if $f(v|A) \geq f(v|B)$ for all $A \subseteq B \subseteq V \setminus \{v\}$, $v \in V$.

Example Submodular: Number of Colors of Balls in Urns

- Consider an urn containing colored balls. Given a set S of balls, $f(S)$ counts the number of distinct colors in S.



Initial value: 2 (colors in urn).
New value with added blue ball: 3

Initial value: 3 (colors in urn).
New value with added blue ball: 3

- Submodularity: Incremental Value of Object Diminishes in a Larger Context (diminishing returns).
- Thus, f is submodular.

Two Equivalent Supermodular Definitions

Definition 2.2.1 (supermodular)

A function $f : 2^V \to \mathbb{R}$ is supermodular if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \leq f(A \cup B) + f(A \cap B) \tag{2.7}$$

Definition 2.2.2 (supermodular (improving returns))

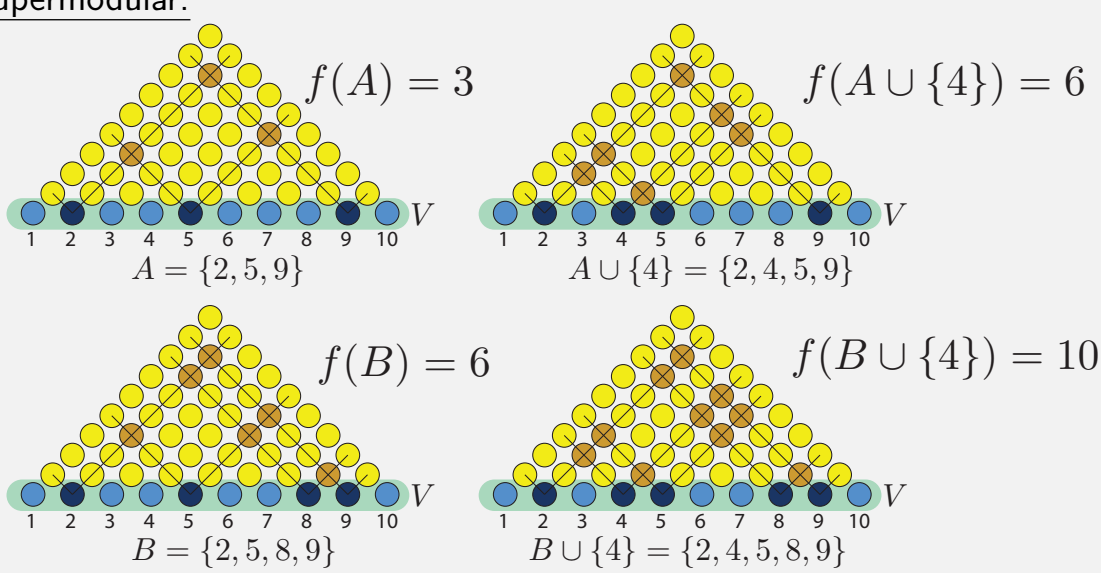
A function $f : 2^V \to \mathbb{R}$ is supermodular if for any $A \subseteq B \subset V$, and $v \in V \setminus B$, we have that:

$$f(A \cup \{v\}) - f(A) \leq f(B \cup \{v\}) - f(B) \tag{2.8}$$

- Incremental "value", "gain", or "cost" of v increases (improves) as the context in which v is considered grows from A to B.
- A function f is submodular iff $-f$ is supermodular.
- If f both submodular and supermodular, then f is said to be **modular**, and $f(A) = c + \sum_{a \in A} \bar{f}(a)$ for some $\bar{f}$ (often $c = 0$).

Example Supermodular: Number of Balls with Two Lines

Given ball pyramid, bottom row V is size $n = |V|$. For subset $S \subseteq V$ of bottom-row balls, draw 45° and 135° diagonal lines from each $s \in S$. Let $f(S)$ be number of non-bottom-row balls with two lines $\Rightarrow f(S)$ is supermodular.

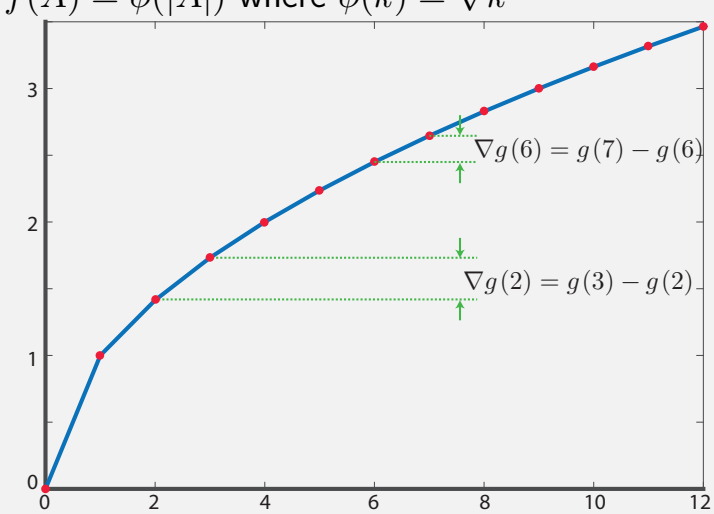


Submodularity's utility in ML

- A model of a physical process:
  - When maximizing, submodularity naturally models: diversity, coverage, span, and information.
  - When minimizing, submodularity naturally models: cooperative costs, complexity, roughness, and irregularity.
  - vice-versa for supermodularity.
- A submodular function can act as a parameter for a machine learning strategy (active/semi-supervised learning, discrete divergence, structured sparse convex norms for use in regularization).
- Itself, as an object or function to learn, based on data.
- A surrogate or relaxation strategy for optimization or analysis
  - An alternate to factorization, decomposition, or sum-product based simplification (as one typically finds in a graphical model). I.e., a means towards tractable surrogates for graphical models.
  - Also, we can "relax" a problem to a submodular one where it can be efficiently solved and offer a bounded quality solution.
  - Non-submodular problems can be analyzed via submodularity.

Square root of cardinality

- Consider $f(A) = \phi(|A|)$ where $\phi(k) = \sqrt{k}$



- Hence, $f(v|A) = \sqrt{|A|+1} - \sqrt{|A|} = \sqrt{k+1} - \sqrt{k}$ when $|A| = k$ and is decreasing in k.
- Hence f is submodular.
- Note that ϕ is a concave function.

Log of modular

- Given a non-negative normalized modular function $m(A) = \sum_{a \in A} m(a)$ where $m(a) \geq 0$.
- Consider $f(A) = \phi(m(A))$ where $\phi(x) = \log(1 + x)$ for $x \in \mathbb{R}_+$.
- Then for $A \subseteq V$,

$$f(v|A) = \log(1 + m(v) + \sum_{a \in A} m(a)) - \log(1 + \sum_{a \in A} m(a)) \tag{2.1}$$

$$= \log(\frac{1 + m(v) + m(A)}{1 + m(A)}) \tag{2.2}$$

- Since $f(v|A) \geq f(v|B)$, f is submodular.

Sums of concave composed with modular

- Let U be a set of indices.
- For for $u \in U$, and $X \subseteq V$, define $m_u(X) = \sum_{x \in X} m_u(x)$, so $m_u(X)$ is a non-negative modular function for all u.
- Let ϕ_u be concave, then $f_u(X) = \phi_u(m_u(X))$ is a submodular function.
- Consider the following class of functions

$$f(X) = \sum_{u \in U} \alpha_u \phi_u(m_u(X)) \tag{2.3}$$

  where ϕ_u is concave, and $\alpha_u \geq 0$ is a non-negative importance weight. It can be shown that f is submodular. This class is known as sums of concave composed with modular (SCCM).
- If all concave functions are non-negative monotone-nondecreasing, we can compose this with another concave function ϕ to reach:

$$f(X) = \phi(\sum_{u \in U} \alpha_u \phi_u(m_u(X))) \tag{2.4}$$

- This is a simple example of a deep submodular function

Variable/Feature Selection in Classification/Regression

- Let Y be a random variable we wish to accurately predict based on at most $n = |V|$ observed measurement variables $(X_1, X_2, \dots, X_n) = X_V$ in a probability model $\Pr(Y, X_1, X_2, \dots, X_n)$.
- Too costly to use all V variables. Goal: choose subset $A \subseteq V$ of variables within budget $|A| \leq k$. Predictions based on only $\Pr(y|x_A)$, hence subset A should retain accuracy.
- The mutual information function $f(A) = I(Y; X_A)$ ("information gain") measures how well variables A can predicting Y (entropy reduction, reduction of uncertainty of Y).
- The mutual information function $f(A) = I(Y; X_A)$ is defined as:

$$I(Y; X_A) = \sum_{y, x_A} \Pr(y, x_A) \log \frac{\Pr(y, x_A)}{\Pr(y)\Pr(x_A)} = H(Y) - H(Y|X_A) \tag{2.5}$$

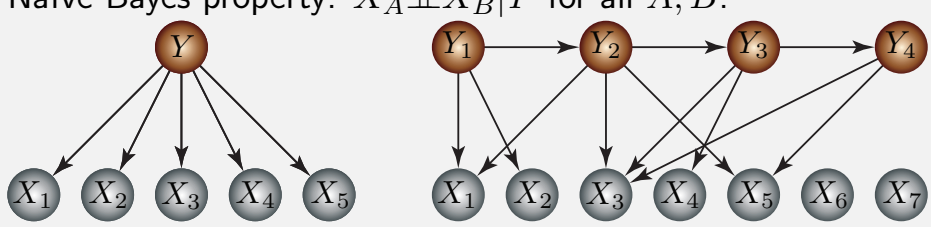
$$= H(X_A) - H(X_A|Y) = H(X_A) + H(Y) - H(X_A, Y) \tag{2.6}$$

- Applicable in pattern recognition, also in sensor coverage problem, where Y is whatever question we wish to ask about environment.

Information Gain and Feature Selection

in Pattern Classification: Naïve Bayes

- Naïve Bayes property: $X_A \perp\!\!\!\perp X_B | Y$ for all A, B.



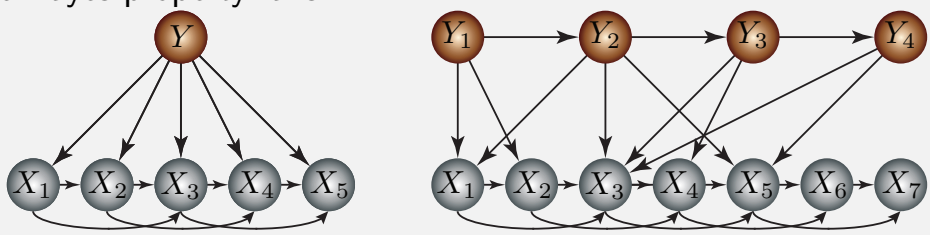
- When $X_A \perp\!\!\!\perp X_B | Y$ for all A, B (the Naïve Bayes assumption holds), then

$$f(A) = I(Y; X_A) = H(X_A) - H(X_A|Y) = H(X_A) - \sum_{a \in A} H(X_a|Y) \tag{2.7}$$

is submodular (submodular minus modular).

Variable Selection in Pattern Classification

- Naïve Bayes property fails:



- $f(A)$ naturally expressed as a difference of two submodular functions

$$f(A) = I(Y; X_A) = H(X_A) - H(X_A|Y), \tag{2.8}$$

  which is a DS (difference of submodular) function.
- Alternatively, when Naïve Bayes assumption is false, we can make a submodular approximation (Peng-2005). E.g., functions of the form:

$$f(A) = \sum_{a \in A} I(X_a; Y) - \lambda \sum_{a, a' \in A} I(X_a; X_{a'}|Y) \tag{2.9}$$

  where $\lambda \geq 0$ is a tradeoff constant.

Variable Selection: Linear Regression Case

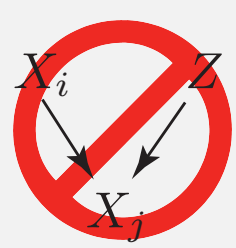
- Next, let Z be continuous. Predictor is linear $\tilde{Z}_A = \sum_{i \in A} \alpha_i X_i$.
- Goodness measure is the squared multiple correlation, i.e.,

$$R^2_{Z,A} = \frac{\text{Var}(Z) - E[(Z - \tilde{Z}_A)^2]}{\text{Var}(Z)} \tag{2.10}$$

  we wish to find A of a given size that maximizes $R^2_{Z,A}$.
- $R^2_{Z,A}$'s maximizing parameters (i.e., under parameters for best linear predictor), for a given A, can be analytically computed. $R^2_{Z,A} = b_A{}^\intercal (C_A^{-1})^\intercal b_A$ whenever $\text{Var}(Z) = 1$, where $b_i = \text{Cov}(Z, X_i)$ and $C = E[(X - E[X])^\intercal (X - E[X])]$ is the covariance matrix.
- When $\exists$ no "suppressor" variables (no v-structures that converge on X_j with parents X_i and Z), then

$$f(A) = R^2_{Z,A} = b_A{}^\intercal (C_A^{-1})^\intercal b_A \tag{2.11}$$

  is a submodular function (so the greedy algorithm gives the $1 - 1/e$ guarantee). (Das&Kempe).



Data/Feature Subset Selection

- Suppose we are given a large data set $\mathcal{D} = \{x_i\}_{i=1}^n$ of n data items $V = \{v_1, v_2, \ldots, v_n\}$ and we wish to choose a subset $A \subset V$ of items that is good in some way (e.g., a summary, or coreset, or sketch).
- Suppose moreover each data item $v \in V$ is described by a vector of non-negative scores for a set U of features (or "properties", or "concepts", etc.) of each data item.
- That is, for $u \in U$ and $v \in V$, let $m_u(v)$ represent the "degree of u-ness" possessed by data item v. Then $m_u \in \mathbb{R}_+^V$ for all $u \in U$.
- Example: U could be a set of colors, and for an image $v \in V$, $m_u(v)$ could represent the number of pixels that are of color u.
- Example: U might be a set of textual features (e.g., ngrams), and $m_u(v)$ is the number of ngrams of type u in sentence v. E.g., if a document consists of the sentence

  $v =$ "Whenever I go to New York City, I visit the New York City museum."

  then $m_{\text{'the'}}(v) = 1$ while $m_{\text{'New York City'}}(v) = 2$.

Data Subset Selection

- For $X \subseteq V$, define $m_u(X) = \sum_{x \in X} m_u(x)$, so $m_u(X)$ is a modular function representing the "degree of u-ness" in subset X.
- Since $m_u(X)$ is modular, it does not have a diminishing returns property. I.e., as we add to X, the degree of u-ness grows additively.
- With ϕ non-decreasing concave, $\phi(m_u(X))$ grows subadditively (if we add v to a context A with less u-ness, the u-ness benefit is more than if we add v to a context $B \supseteq A$ having more u-ness). That is

$$\phi(m_u(A + v)) - \phi(m_u(A)) \geq \phi(m_u(B + v)) - \phi(m_u(B)) \tag{2.12}$$

- Consider the following class of feature functions $f : 2^V \to \mathbb{R}_+$

$$f(X) = \sum_{u \in U} \alpha_u \phi_u(m_u(X)) \tag{2.13}$$

  where ϕ_u is a non-decreasing concave, and $\alpha_u \geq 0$ is a feature importance weight. Thus, f is submodular.
- $f(X)$ measures X's ability to represent set of features U as measured by $m_u(X)$, with diminishing returns function ϕ, and importance weights α_u.

Data Subset Selection, KL-divergence

- Let $p = \{p_u\}_{u \in U}$ be a desired probability distribution over features (i.e., $\sum_u p_u = 1$ and $p_u \geq 0$ for all $u \in U$).
- Next, normalize the modular weights for each feature:

$$0 \leq \bar{m}_u(X) \triangleq \frac{m_u(X)}{\sum_{u' \in U} m_{u'}(X)} = \frac{m_u(X)}{m(X)} \leq 1 \tag{2.14}$$

  where $m(X) \triangleq \sum_{u' \in U} m_{u'}(X)$.
- Then for any $X \subseteq V$, $\bar{m}_u(X)$ can also be seen as a distribution over features U since $\bar{m}_u(X) \geq 0$ and $\sum_{u \in U} \bar{m}_u(X) = 1$.
- Consider the KL-divergence between these two distributions:

$$D(p || \{\bar{m}_u(X)\}_{u \in U}) = \sum_{u \in U} p_u \log p_u - \sum_{u \in U} p_u \log(\bar{m}_u(X)) \tag{2.15}$$

$$= \sum_{u \in U} p_u \log p_u - \sum_{u \in U} p_u \log(m_u(X)) + \log(m(X))$$

$$= -H(p) + \log m(X) - \sum_{u \in U} p_u \log(m_u(X)) \tag{2.16}$$

Data Subset Selection, KL-divergence

- The objective once again, treating entropy $H(p)$ as a constant,

$$D(p || \{\bar{m}_u(X)\}) = \text{const.} + \log m(X) - \sum_{u \in U} p_u \log(m_u(X)) \tag{2.17}$$

- But seen as a function of X, both $\log m(X)$ and $\sum_{u \in U} p_u \log m_u(X)$ are submodular functions.
- Hence the KL-divergence, seen as a function of X, i.e., $f(X) = D(p || \{\bar{m}_u(X)\})$ is quite naturally represented as a difference of submodular functions.
- Alternatively, if we define (Shinohara, 2014)

$$g(X) \triangleq \log m(X) - D(p || \{\bar{m}_u(X)\}) = \sum_{u \in U} p_u \log(m_u(X)) \tag{2.18}$$

  we have a submodular function g that represents a combination of its quantity of X via its features (i.e., $\log m(X)$) and its feature distribution closeness to some distribution p (i.e., $D(p || \{\bar{m}_u(X)\})$).

Other examples as diversity models

- Sensor placement
- Social networks and influential nodes
- Viral marketing, information cascades, diffusion networks

Recall from lecture 1

- Next slide comes from lecture 1.

Sets, Characteristic Vectors, and pseudo-Boolean functions

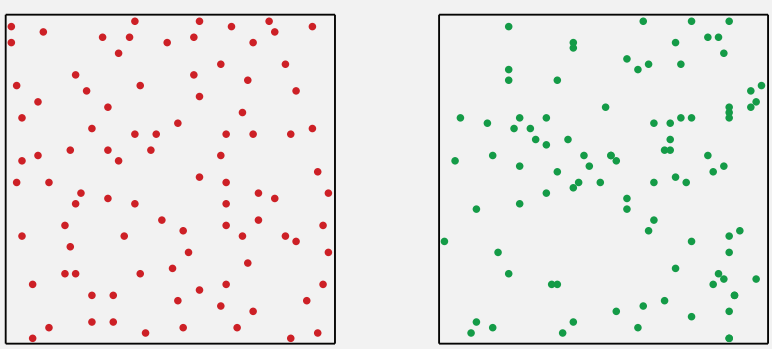
- Any set $A \subseteq V$ can be represented as a binary vector $x \in \{0,1\}^V$ (a "bit vector" representation of a set).
- The characteristic vector $\mathbf{1}_A \in \{0,1\}^V$ of a set A is defined one where element $v \in V$ has value:

$$\mathbf{1}_A(v) = \begin{cases} 1 & \text{if } v \in A \\ 0 & else \end{cases} \tag{2.7}$$

- Useful to be able to quickly map from set to binary vector, 1_A, and from vector to set, $A = V(\mathbf{1}_A)$.
- Given $x \in \{0,1\}^V$, map to set via $V(x) \subseteq V$, where $v \in V(x)$ iff $x(v) = 1$.
- $f : \{0,1\}^V \to \{0,1\}$ are known as Boolean function.
- $f : \{0,1\}^V \to \mathbb{R}$ is a pseudo-Boolean function (submodular functions are a special case).

Determinantal Point Processes (DPPs)

- Sometimes we wish not only to valuate subsets $A \subseteq V$ but to induce probability distributions over all subsets.
- We may wish to prefer samples where elements of A are diverse (i.e., given a sample A, for $a, b \in A$, we prefer a and b to be different).



DPP Independent

(Kulesza, Gillenwater, & Taskar, 2011)

- A Determinantal point processes (DPPs) is a probability distribution over subsets A of V where the "energy" function is submodular.
- More "diverse" or "complex" samples are given higher probability.

DPPs and log-submodular probability distributions

- Given binary vectors $x, y \in \{0,1\}^V$, $y \leq x$ if $y(v) \leq x(v), \forall v \in V$.
- Given a positive-definite $n \times n$ matrix M, a subset $X \subseteq V$, let M_X be $|X| \times |X|$ principle submatrix, rows/columns specified by $X \subseteq V$.
- A Determinantal Point Process (DPP) is a distribution of the form:

$$\Pr(\mathbf{X} = x) = \frac{|M_{V(x)}|}{|M+I|} = \exp\left(\log\left(\frac{|M_{V(x)}|}{|M+I|}\right)\right) \propto \det(M_{V(x)}) \tag{2.19}$$

  where I is $n \times n$ identity matrix, and $\mathbf{X} \in \{0,1\}^V$ is a random vector.
- Equivalently, defining K as $K = M(M+I)^{-1}$, we have:

$$\sum_{x \in \{0,1\}^V : x \geq y} \Pr(\mathbf{X} = x) = \Pr(\mathbf{X} \geq y) = \exp\left(\log\left(|K_{V(y)}|\right)\right) \tag{2.20}$$

- Given positive definite matrix M, function $f : 2^V \to \mathbb{R}$ with $f(A) = \log |M_A|$ (the logdet function) is submodular.
- Therefore, <u>a DPP is a log-submodular probability distribution</u>.

Graphical Models and fast MAP Inference

- Distribution over $x \in \{0,1\}^V$ that factors w.r.t. a graph:

$$p(x) = \frac{1}{Z} \exp(-E(x)) \tag{2.21}$$

  where $E(x) = \sum_{c \in \mathcal{C}} E_c(x_c)$ and $\mathcal{C}$ are cliques of graph $G = (V, \mathcal{E})$.
- <u>MAP inference</u> problem is important in ML: compute

$$x^* \in \operatorname*{argmax}_{x \in \{0,1\}^V} p(x) \tag{2.22}$$

- Easy when G a tree, exponential in k (tree-width of G) in general.
- Even worse, NP-hard to find the tree-width.
- Tree-width can be large even when degree is small (e.g., regular grid graphs have low-degree but $\Omega(\sqrt{n})$ tree-width).
- Many approximate inference strategies utilize additional factorization assumptions (e.g., mean-field, variational inference, expectation propagation, etc).
- Can we do exact MAP inference in polynomial time regardless of the tree-width, without even knowing the tree-width? Yes!

Submodular Generalized Dependence

- there is a notion of "independence" , i.e., $A \perp\!\!\!\perp B$:

$$f(A \cup B) = f(A) + f(B), \tag{2.23}$$

- and a notion of "conditional independence" , i.e., $A \perp\!\!\!\perp B | C$:

$$f(A \cup B \cup C) + f(C) = f(A \cup C) + f(B \cup C) \tag{2.24}$$

- and a notion of "dependence" (conditioning reduces valuation):

$$f(A|B) \triangleq f(A \cup B) - f(B) < f(A), \tag{2.25}$$

- and a notion of "conditional mutual information"

$$I_f(A; B|C) \triangleq f(A \cup C) + f(B \cup C) - f(A \cup B \cup C) - f(C) \geq 0$$

- and two notions of "information amongst a collection of sets":

$$I_f(S_1; S_2; \dots; S_k) = \sum_{i=1}^{k} f(S_k) - f(S_1 \cup S_2 \cup \dots \cup S_k) \tag{2.26}$$

$$I'_f(S_1; S_2; \dots; S_k) = \sum_{A \subseteq \{1,2,\dots,k\}} (-1)^{|A|+1} f(\bigcup_{j \in A} S_j) \tag{2.27}$$

Submodular Parameterized Clustering

- Given a submodular function $f : 2^V \to \mathbb{R}$, form the combinatorial dependence function $I_f(A; B) = f(A) + f(B) - f(A \cup B)$.
- Consider clustering algorithm: First find partition $A_1^* \in \operatorname{argmin}_{A \subseteq V} I_f(A; V \setminus A)$ and $A_2^* = V \setminus A_1^*$.
- Then partition the partitions: $A_{11}^* \in \operatorname{argmin}_{A \subseteq A_1^*} I_f(A; A_1^* \setminus A)$, $A_{12}^* = A_1^* \setminus A_{11}^*$, and $A_{21}^* \in \operatorname{argmin}_{A \subseteq A_2^*} I_f(A; A_2^* \setminus A)$, etc.
- Recursively partition the partitions, we end up with a partition $V = V_1 \cup V_2 \cup \dots \cup V_k$ that clusters the data.
- Each minimization can be done using Queyranne's algorithm (alternatively can construct a Gomory-Hu tree). This gives a partition no worse than factor 2 away from optimal partition. (Narasimhan&Bilmes, 2007).
- Hence, family of clustering algorithms parameterized by f.

Ground set: E or V?

Submodular functions are functions defined on subsets of some finite set, called the ground set.

- It is common in the literature to use either E or V as the ground set — we will at different times use both (there should be no confusion).
- The terminology ground set comes from lattice theory, where V are the ground elements of a lattice (just above 0).

Notation $\mathbb{R}^E$, and modular functions as vectors

What does $x \in \mathbb{R}^E$ mean?

$$\mathbb{R}^E = \{x = (x_j \in \mathbb{R} : j \in E)\} \tag{2.28}$$

and

$$\mathbb{R}^E_+ = \{x = (x_j : j \in E) : x \geq 0\} \tag{2.29}$$

Any vector $x \in \mathbb{R}^E$ can be treated as a normalized modular function, and vice verse. That is, for $A \subseteq E$,

$$x(A) = \sum_{a \in A} x_a \tag{2.30}$$

Note that x is said to be normalized since $x(\emptyset) = 0$.

Characteristic vectors of sets & modular functions

- Given an $A \subseteq E$, define the incidence (or characteristic) vector $\mathbf{1}_A \in \{0,1\}^E$ on the unit hypercube to be

$$\mathbf{1}_A(j) = \begin{cases} 1 & \text{if } j \in A; \\ 0 & \text{if } j \notin A \end{cases} \tag{2.31}$$

  or equivalently,

$$\mathbf{1}_A \stackrel{\text{def}}{=} \left\{ x \in \{0,1\}^E : x_i = 1 \text{ iff } i \in A \right\} \tag{2.32}$$

- Sometimes this is written as $\chi_A \equiv \mathbf{1}_A$.
- Thus, given modular function $x \in \mathbb{R}^E$, we can write $x(A)$ in a variety of ways, i.e.,

$$x(A) = x^\intercal \cdot \mathbf{1}_A = \langle x, \mathbf{1}_A \rangle = \sum_{i \in A} x(i) \tag{2.33}$$

Other Notation: singletons and sets

When A is a set and k is a singleton (i.e., a single item), the union is properly written as $A \cup \{k\}$, but sometimes we will write just $A + k$.

What does S^T mean when S and T are arbitrary sets?

- Let S and T be two arbitrary sets (either of which could be countable, or uncountable).
- We define the notation S^T to be the set of all functions that map from T to S. That is, if $f \in S^T$, then $f : T \to S$.
- Hence, given a finite set E, $\mathbb{R}^E$ is the set of all functions that map from elements of E to the reals $\mathbb{R}$, and such functions are identical to a vector in a vector space with axes labeled as elements of E (i.e., if $m \in \mathbb{R}^E$, then for all $e \in E$, $m(e) \in \mathbb{R}$).
- Often "2" is shorthand for the set $\{0, 1\}$. I.e., $\mathbb{R}^2$ where $2 \equiv \{0, 1\}$.
- Similarly, 2^E is the set of all functions from E to "two" — so 2^E is shorthand for $\{0, 1\}^E$ — hence, 2^E is the set of all functions that map from elements of E to $\{0, 1\}$, equivalent to all binary vectors with elements indexed by elements of E, equivalent to subsets of E. Hence, if $A \in 2^E$ then $A \subseteq E$.
- What might 3^E mean?

Example Submodular: Entropy from Information Theory

- Entropy is submodular. Let V be the index set of a set of random variables, then the function

$$f(A) = H(X_A) = -\sum_{x_A} p(x_A) \log p(x_A) \tag{2.34}$$

is submodular.

- Proof: (further) conditioning reduces entropy. With $A \subseteq B$ and $v \notin B$,

$$H(X_v|X_B) = H(X_{B+v}) - H(X_B) \tag{2.35}$$
$$\leq H(X_{A+v}) - H(X_A) = H(X_v|X_A) \tag{2.36}$$

- We say "further" due to $B \setminus A$ not nec. empty.

Example Submodular: Entropy from Information Theory

- Alternate Proof: Conditional mutual Information is always non-negative.
- Given $A, B \subseteq V$, consider conditional mutual information quantity:

$$I(X_{A\setminus B}; X_{B\setminus A}|X_{A\cap B}) = \sum_{x_{A\cup B}} p(x_{A\cup B}) \log \frac{p(x_{A\setminus B}, x_{B\setminus A}|x_{A\cap B})}{p(x_{A\setminus B}|x_{A\cap B})p(x_{B\setminus A}|x_{A\cap B})}$$
$$= \sum_{x_{A\cup B}} p(x_{A\cup B}) \log \frac{p(x_{A\cup B})p(x_{A\cap B})}{p(x_A)p(x_B)} \geq 0 \quad (2.37)$$

then

$$I(X_{A\setminus B}; X_{B\setminus A}|X_{A\cap B}) = H(X_A) + H(X_B) - H(X_{A\cup B}) - H(X_{A\cap B}) \geq 0 \quad (2.38)$$

so entropy satisfies

$$H(X_A) + H(X_B) \geq H(X_{A\cup B}) + H(X_{A\cap B}) \quad (2.39)$$

Information Theory: Block Coding

- Given a set of random variables $\{X_i\}_{i\in V}$ indexed by set V, how do we partition them so that we can best block-code them within each block.
- I.e., how do we form $S \subseteq V$ such that $I(X_S; X_{V\setminus S})$ is as small as possible, where $I(X_A; X_B)$ is the mutual information between random variables X_A and X_B, i.e.,

$$I(X_A; X_B) = H(X_A) + H(X_B) - H(X_A, X_B) \quad (2.40)$$

and $H(X_A) = -\sum_{x_A} p(x_A) \log p(x_A)$ is the joint entropy of the set X_A of random variables.

Example Submodular: Mutual Information

- Also, symmetric mutual information is submodular,

$$f(A) = I(X_A; X_{V \setminus A}) = H(X_A) + H(X_{V \setminus A}) - H(X_V) \tag{2.41}$$

Note that $f(A) = H(X_A)$ and $\bar{f}(A) = H(X_{V \setminus A})$, and adding submodular functions preserves submodularity (which we will see quite soon).

Monge Matrices

- $m \times n$ matrices $C = [c_{ij}]_{ij}$ are called Monge matrices if they satisfy the Monge property, namely:

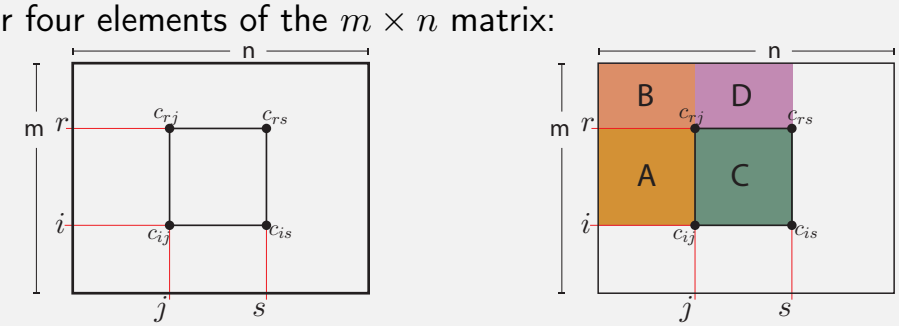
$$c_{ij} + c_{rs} \leq c_{is} + c_{rj} \tag{2.42}$$

for all $1 \leq i < r \leq m$ and $1 \leq j < s \leq n$.

- Equivalently, for all $1 \leq i, r \leq m$, $1 \leq j, s \leq n$,

$$c_{\min(i,r),\min(j,s)} + c_{\max(i,r),\max(j,s)} \leq c_{is} + c_{rj} \tag{2.43}$$

- Consider four elements of the $m \times n$ matrix:



$c_{ij} = A + B$, $c_{rj} = B$, $c_{rs} = B + D$, $c_{is} = A + B + C + D$.

Monge Matrices, where useful

- Useful for speeding up transportation, dynamic programming, flow, search, lot-sizing and many other problems.
- Example, Hitchcock transportation problem: Given $m \times n$ cost matrix $C = [c_{ij}]_{ij}$, a non-negative supply vector $a \in \mathbb{R}_+^m$, a non-negative demand vector $b \in \mathbb{R}_+^n$ with $\sum_{i=1}^m a(i) = \sum_{j=1}^n b_j$, we wish to optimally solve the following linear program:

$$\underset{X \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij} \tag{2.44}$$

$$\text{subject to} \quad \sum_{i=1}^m x_{ij} = b_j \;\; \forall j = 1, \dots, n \tag{2.45}$$

$$\sum_{j=1}^n x_{ij} = a_i \;\; \forall i = 1, \dots, m \tag{2.46}$$

$$x_{i,j} \geq 0 \;\; \forall i, j \tag{2.47}$$

Monge Matrices, Hitchcock transportation

C

Producers, Sources, or Supply

a_1	2	0	1	3	3
a_2	1	1	4	7	10
a_3	5	0	4	9	14
		3	2	1	2
		b_1	b_2	b_3	b_4

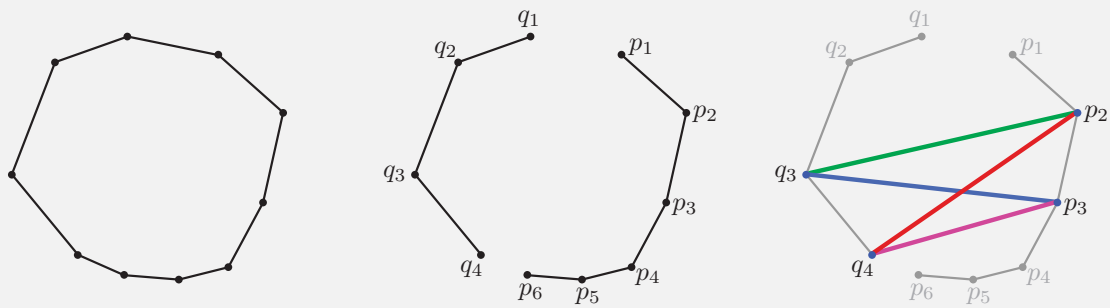
Consumers, Sinks, or Demand

- Solving the linear program can be done easily and optimally using the "North-West Corner Rule" (a 2D greedy-like approach starting at top-left and moving down-right) in only $O(m + n)$ if the matrix C is Monge!

Monge Matrices and Convex Polygons

- Can generate a Monge matrix from a convex polygon - delete two segments, then separately number vertices on each chain. Distances c_{ij} satisfy Monge property (or quadrangle inequality).

$$d(p_2, q_3) + d(p_3, q_4) \leq d(p_2, q_4) + d(p_3, q_3) \tag{2.48}$$



Monge Matrices and Submodularity

- A submodular function has the form: $f : 2^V \to \mathbb{R}$ which can be seen as $f : \{0,1\}^V \to \mathbb{R}$
- We can generalize this to $f : \{0, 1, \ldots, K\}^V \to \mathbb{R}$ for some constant $K \in \mathbb{Z}_+$.
- We may define submodularity as: for all $x, y \in \{0, 1, \ldots, K\}^V$, we have

$$f(x) + f(y) \geq f(x \vee y) + f(x \wedge y) \tag{2.49}$$

- $x \vee y$ is the (join) element-wise max of each element, that is $(x \vee y)(v) = \max(x(v), y(v))$ for $v \in V$.
- $x \wedge y$ is the (meet) element-wise min of each element, that is, $(x \wedge y)(v) = \min(x(v), y(v))$ for $v \in V$.
- With $K = 1$, then this is the standard definition of submodularity.
- With $|V| = 2$, and $K + 1$ the side-dimension of the matrix, we get a Monge property (on square matrices).
- Non square: $f : \{0, 1, \ldots, K_1\} \times \{0, 1, \ldots, K_2\} \to \mathbb{R}$.

Submodular Motivation Recap

- Given a set of objects $V = \{v_1, \ldots, v_n\}$ and a function $f : 2^V \to \mathbb{R}$ that returns a real value for any subset $S \subseteq V$.
- Suppose we are interested in finding the subset that either maximizes or minimizes the function, e.g., $\operatorname{argmax}_{S \subseteq V} f(S)$, possibly subject to some constraints.
- In general, this problem has exponential time complexity.
- Example: f might correspond to the value (e.g., information gain) of a set of sensor locations in an environment, and we wish to find the best set $S \subseteq V$ of sensors locations given a fixed upper limit on the number of sensors $|S|$.
- In many cases (such as above) f has properties that make its optimization tractable to either exactly or approximately compute.
- One such property is *submodularity*.

Two Equivalent Submodular Definitions

Definition 2.10.1 (submodular concave)

A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \tag{2.7}$$

An alternate and (as we will soon see) equivalent definition is:

Definition 2.10.2 (diminishing returns)

A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A \subseteq B \subset V$, and $v \in V \setminus B$, we have that:

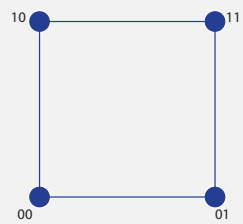
$$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B) \tag{2.8}$$

• The incremental "value", "gain", or "cost" of v decreases (diminishes) as the context in which v is considered grows from A to B.
• Gain notation: Define $f(v|A) \triangleq f(A + v) - f(A)$. Then function f is submodular if $f(v|A) \geq f(v|B)$ for all $A \subseteq B \subseteq V \setminus \{v\}$, $v \in V$.

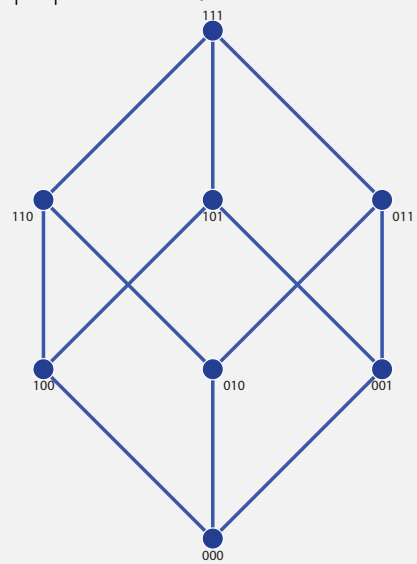
Submodular on Hypercube Vertices

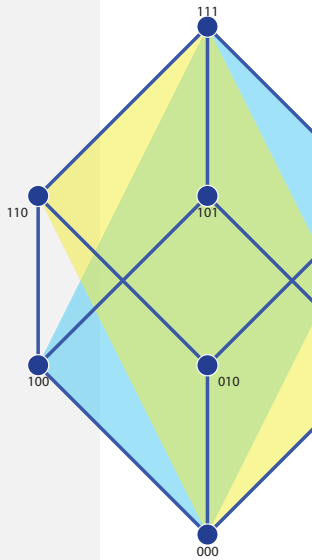
We can test submodularity via values on vertices of hypercube.

Example: with $|V| = n = 2$, this is easy:



With $|V| = n = 3$, a bit harder.





How many inequalities of form $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$?

Subadditive Definitions

Definition 2.10.1 (subadditive)

A function $f : 2^V \to \mathbb{R}$ is subadditive if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \geq f(A \cup B) \tag{2.50}$$

This means that the "whole" is less than the sum of the parts.

Two Equivalent Supermodular Definitions

Definition 2.10.1 (supermodular)

A function $f : 2^V \to \mathbb{R}$ is supermodular if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \leq f(A \cup B) + f(A \cap B) \tag{2.7}$$

Definition 2.10.2 (supermodular (improving returns))

A function $f : 2^V \to \mathbb{R}$ is supermodular if for any $A \subseteq B \subset V$, and $v \in V \setminus B$, we have that:

$$f(A \cup \{v\}) - f(A) \leq f(B \cup \{v\}) - f(B) \tag{2.8}$$

- Incremental "value", "gain", or "cost" of v increases (improves) as the context in which v is considered grows from A to B.
- A function f is submodular iff $-f$ is supermodular.
- If f both submodular and supermodular, then f is said to be **modular**, and $f(A) = c + \sum_{a \in A} \bar{f}(a)$ for some $\bar{f}$ (often $c = 0$).

Superadditive Definitions

Definition 2.10.2 (superadditive)

A function $f : 2^V \to \mathbb{R}$ is superadditive if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \leq f(A \cup B) \tag{2.51}$$

- This means that the "whole" is greater than the sum of the parts.
- In general, submodular and subadditive (and supermodular and superadditive) are different properties.
- Ex: Let $0 < k < |V|$, and consider $f : 2^V \to \mathbb{R}_+$ where:

$$f(A) = \begin{cases} 1 & \text{if } |A| \leq k \\ 0 & \text{else} \end{cases} \tag{2.52}$$

- This function is subadditive but not submodular.

Modular Definitions

Definition 2.10.3 (modular)

A function that is both submodular and supermodular is called **modular**

If f is a modular function, than for any $A, B \subseteq V$, we have

$$f(A) + f(B) = f(A \cap B) + f(A \cup B) \tag{2.53}$$

In modular functions, elements do not interact (or cooperate, or compete, or influence each other), and have value based only on singleton values.

Proposition 2.10.4

If f is modular, it may be written as

$$f(A) = f(\emptyset) + \sum_{a \in A} \Big(f(\{a\}) - f(\emptyset) \Big) = c + \sum_{a \in A} f'(a) \tag{2.54}$$

which has only $|V| + 1$ parameters.

Modular Definitions

Proof.

We inductively construct the value for $A = \{a_1, a_2, \ldots, a_k\}$.
For $k = 2$,

$$f(a_1) + f(a_2) = f(a_1, a_2) + f(\emptyset) \tag{2.55}$$

$$\text{implies } f(a_1, a_2) = f(a_1) - f(\emptyset) + f(a_2) - f(\emptyset) + f(\emptyset) \tag{2.56}$$

then for $k = 3$,

$$f(a_1, a_2) + f(a_3) = f(a_1, a_2, a_3) + f(\emptyset) \tag{2.57}$$

$$\text{implies } f(a_1, a_2, a_3) = f(a_1, a_2) - f(\emptyset) + f(a_3) - f(\emptyset) + f(\emptyset) \tag{2.58}$$

$$= f(\emptyset) + \sum_{i=1}^{3} \big(f(a_i) - f(\emptyset) \big) \tag{2.59}$$

and so on ... □

Complement function

Given a function $f : 2^V \to \mathbb{R}$, we can find a complement function $\bar{f} : 2^V \to \mathbb{R}$ as $\bar{f}(A) = f(V \setminus A)$ for any A.

Proposition 2.10.5

$\bar{f}$ is submodular iff f is submodular.

Proof.

$$\bar{f}(A) + \bar{f}(B) \geq \bar{f}(A \cup B) + \bar{f}(A \cap B) \tag{2.60}$$

follows from

$$f(V \setminus A) + f(V \setminus B) \geq f(V \setminus (A \cup B)) + f(V \setminus (A \cap B)) \tag{2.61}$$

which is true because $V \setminus (A \cup B) = (V \setminus A) \cap (V \setminus B)$ and $V \setminus (A \cap B) = (V \setminus A) \cup (V \setminus B)$ (De Morgan's laws for sets). □

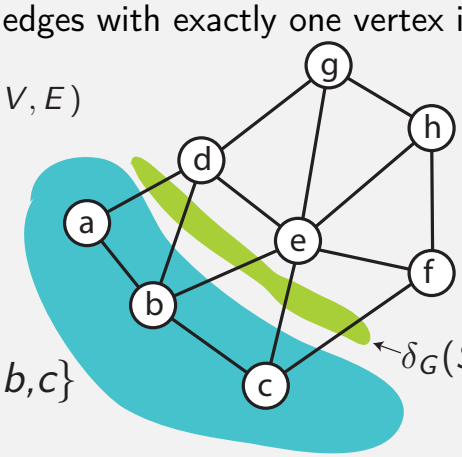
Undirected Graphs

- Let $G = (V, E)$ be a graph with vertices $V = V(G)$ and edges $E = E(G) \subseteq V \times V$.
- If G is undirected, define

$$E(X, Y) = \{\{x, y\} \in E(G) : x \in X \setminus Y, y \in Y \setminus X\} \tag{2.62}$$

  as the edges strictly between X and Y.
- Nodes define cuts. Define the cut function $\delta(X) = E(X, V \setminus X)$, set of edges with exactly one vertex in X.

$G = (V, E)$



$S = \{a, b, c\}$

$\delta_G(S) = \{\{u, v\} \in E : u \in S, v \in V \setminus S\}.$
$= \{\{a,d\},\{b,d\},\{b,e\},\{c,e\},\{c,f\}\}$

Directed graphs, and cuts and flows

- If G is directed, define

$$E^+(X,Y) \triangleq \{(x,y) \in E(G) : x \in X \setminus Y, y \in Y \setminus X\} \tag{2.63}$$

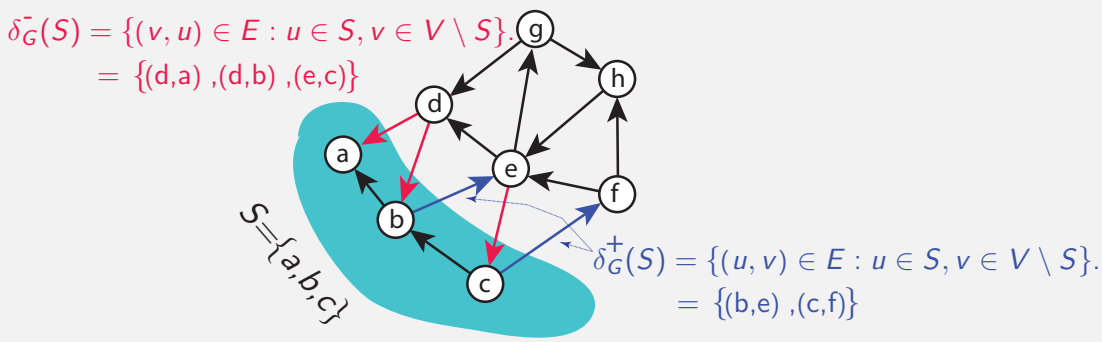
as the edges directed strictly from X towards Y.

- Nodes define cuts and flows. Define edges leaving X (out-flow) as

$$\delta^+(X) \triangleq E^+(X, V \setminus X) \tag{2.64}$$

and edges entering X (in-flow) as

$$\delta^-(X) \triangleq E^+(V \setminus X, X) \tag{2.65}$$

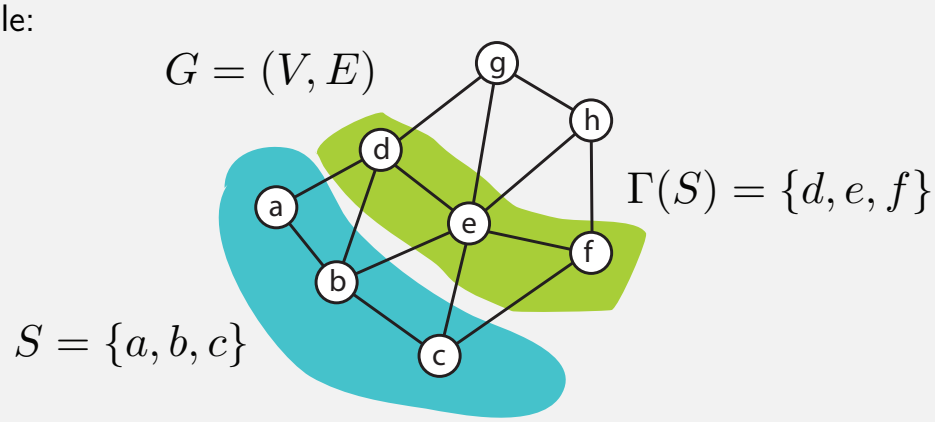


The Neighbor function in undirected graphs

- Given a set $X \subseteq V$, the neighbor function of X is defined as

$$\Gamma(X) \triangleq \{v \in V(G) \setminus X : E(X, \{v\}) \neq \emptyset\} \tag{2.66}$$

- Example:



Directed Cut function: property

Lemma 2.11.1

For a digraph $G = (V, E)$ and any $X, Y \subseteq V$: we have

$$|\delta^+(X)| + |\delta^+(Y)| = |\delta^+(X \cap Y)| + |\delta^+(X \cup Y)| + |E^+(X, Y)| + |E^+(Y, X)| \tag{2.67}$$

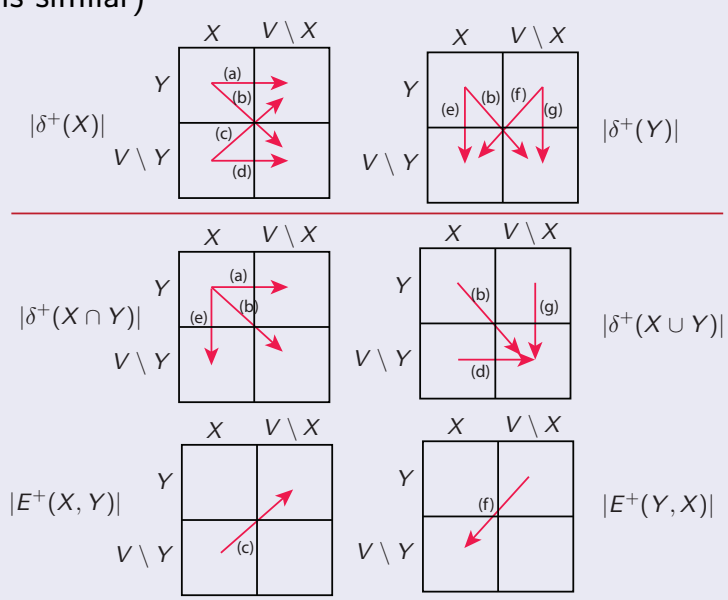
and

$$|\delta^-(X)| + |\delta^-(Y)| = |\delta^-(X \cap Y)| + |\delta^-(X \cup Y)| + |E^-(X, Y)| + |E^-(Y, X)| \tag{2.68}$$

Directed Cut function: proof of property

Proof.

We can prove Eq. (2.67) using a geometric counting argument (proof for $|\delta^-(X)|$ case is similar)



Q:Why is (c) $= |E^+(X, Y)|$?

Directed cut/flow functions: submodular

Lemma 2.11.2

For a digraph $G = (V, E)$ and any $X, Y \subseteq V$: both functions $|\delta^+(X)|$ and $|\delta^-(X)|$ are submodular.

Proof.

$|E^+(X, Y)| \geq 0$ and $|E^-(X, Y)| \geq 0$. □

More generally, in the non-negative weighted edge case, both in-flow and out-flow are submodular on subsets of the vertices.

Undirected Cut/Flow & the Neighbor function: submodular

Lemma 2.11.3

For an undirected graph $G = (V, E)$ and any $X, Y \subseteq V$: we have that both the undirected cut (or flow) function $|\delta(X)|$ and the neighbor function $|\Gamma(X)|$ are submodular. I.e.,

$$|\delta(X)| + |\delta(Y)| = |\delta(X \cap Y)| + |\delta(X \cup Y)| + 2|E(X, Y)| \tag{2.69}$$

and

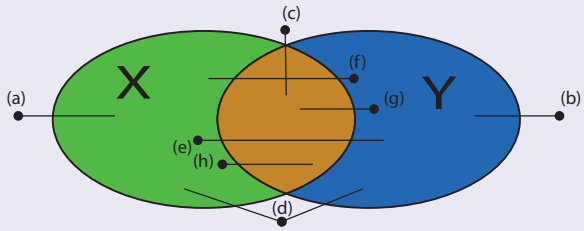
$$|\Gamma(X)| + |\Gamma(Y)| \geq |\Gamma(X \cap Y)| + |\Gamma(X \cup Y)| \tag{2.70}$$

Proof.

- Eq. (2.69) follows from Eq. (2.67): we replace each undirected edge $\{u, v\}$ with two oppositely-directed directed edges (u, v) and (v, u). Then we use same counting argument.
- Eq. (2.70) follows as shown in the following page.

...

cont.



Graphically, we can count and see that

$$\Gamma(X) = (a) + (c) + (f) + (g) + (d) \tag{2.71}$$
$$\Gamma(Y) = (b) + (c) + (e) + (h) + (d) \tag{2.72}$$
$$\Gamma(X \cup Y) = (a) + (b) + (c) + (d) \tag{2.73}$$
$$\Gamma(X \cap Y) = (c) + (g) + (h) \tag{2.74}$$

so

$$\begin{aligned} |\Gamma(X)| + |\Gamma(Y)| &= (a) + (b) + 2(c) + 2(d) + (e) + (f) + (g) + (h) \\ &\geq (a) + (b) + 2(c) + (d) + (g) + (h) = |\Gamma(X \cup Y)| + |\Gamma(X \cap Y)| \end{aligned} \tag{2.75}$$

Undirected Neighbor functions

Therefore, the undirected cut function $|\delta(A)|$ and the neighbor function $|\Gamma(A)|$ of a graph G are both submodular.

Undirected cut/flow is submodular: alternate proof

- Another simple proof shows that $|\delta(X)|$ is submodular.
- Define a graph $G_{uv} = (\{u, v\}, \{e\}, w)$ with two nodes u, v and one edge $e = \{u, v\}$ with non-negative weight $w(e) \in \mathbb{R}_+$.
- Weighted cut function over those two nodes: $w(\delta_{u,v}(\cdot))$ has valuation:

$$w(\delta_{u,v}(\emptyset)) = w(\delta_{u,v}(\{u, v\})) = 0 \tag{2.76}$$

and

$$w(\delta_{u,v}(\{u\})) = w(\delta_{u,v}(\{v\})) = w \geq 0 \tag{2.77}$$

- Thus, $w(\delta_{u,v}(\cdot))$ is submodular since

$$w(\delta_{u,v}(\{u\})) + w(\delta_{u,v}(\{v\})) \geq w(\delta_{u,v}(\{u, v\})) + w(\delta_{u,v}(\emptyset)) \tag{2.78}$$

- General non-negative weighted graph $G = (V, E, w)$, define $w(\delta(\cdot))$:

$$f(X) = w(\delta(X)) = \sum_{(u,v) \in E(G)} w(\delta_{u,v}(X \cap \{u, v\})) \tag{2.79}$$

- This is easily shown to be submodular using properties we will soon see (namely, submodularity closed under summation and restriction).

Other graph functions that are submodular/supermodular

These come from Narayanan's book 1997. Let G be an undirected graph.

- Let $V(X)$ be the vertices adjacent to some edge in $X \subseteq E(G)$, then $|V(X)|$ (the vertex function) is submodular.
- Let $E(S)$ be the edges with both vertices in $S \subseteq V(G)$. Then $|E(S)|$ (the interior edge function) is supermodular.
- Let $I(S)$ be the edges with at least one vertex in $S \subseteq V(G)$. Then $|I(S)|$ (the incidence function) is submodular.
- Recall $|\delta(S)|$, is the number of edges with exactly one vertex in $S \subseteq V(G)$ is submodular (cut function). Thus, we have $I(S) = E(S) \cup \delta(S)$ and $E(S) \cap \delta(S) = \emptyset$, and thus that $|I(S)| = |E(S)| + |\delta(S)|$. So we can get a submodular function by summing a submodular and a supermodular function. If you had to guess, is this always the case?
- Consider $f(A) = |\delta^+(A)| - |\delta^+(V \setminus A)|$. Guess, submodular, supermodular, modular, or neither? Exercise: determine which one and prove it.

Number of connected components in a graph via edges

- Recall, $f : 2^V \to \mathbb{R}$ is submodular, then so is $\bar{f} : 2^V \to \mathbb{R}$ defined as $\bar{f}(S) = f(V \setminus S)$.
- Hence, if $g : 2^V \to \mathbb{R}$ is supermodular, then so is $\bar{g} : 2^V \to \mathbb{R}$ defined as $\bar{g}(S) = g(V \setminus S)$.
- Given a graph $G = (V, E)$, for each $A \subseteq E(G)$, let $c(A)$ denote the number of connected components of the (spanning) subgraph $(V(G), A)$, with $c : 2^E \to \mathbb{R}_+$. Thus, $c(\emptyset) = |V|$, and $c(E) \geq 1$.
- $c(A)$ is monotone non-increasing, $c(A + a) - c(A) \leq 0$.
- Then $c(A)$ is supermodular, i.e.,
$$c(A + a) - c(A) \leq c(B + a) - c(B) \tag{2.80}$$
with $A \subseteq B \subseteq E \setminus \{a\}$.
- Intuition: an edge is "more" (no less) able to bridge separate components (and reduce the number of connected components) when edge is added in a smaller context than when added in a larger context.
- $\bar{c}(A) = c(E \setminus A)$ is number of connected components in G when we remove A; supermodular monotone non-decreasing but not normalized.