CMSC 510 – L14
Regularization Methods for Machine Learning

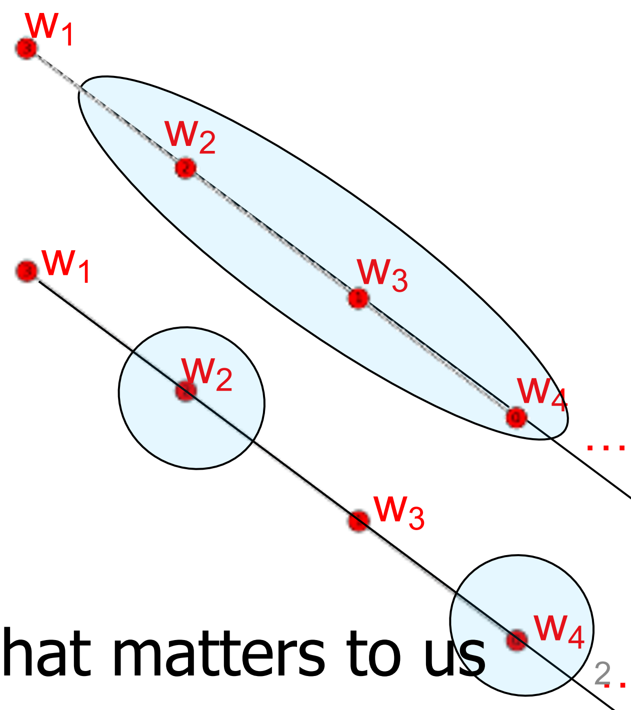
Instructor:
Dr. Tom Arodz

Fused Lasso – view involving sets

- If f_i is important for classification, f_{i-1} and f_{i+1} likely to be important, too
- We want a classifier where either both neighboring features are selected (non-zero w_i and w_{i+1}) or both are not selected (w_i=w_{i+1}=0):

$$\Omega(w) = \sum_{f=2}^{F} \left| [w_{f-1}] - [w_f] \right|$$

$$[x] = \mathrm{supp}(x) = |\mathrm{sign}(x)|$$



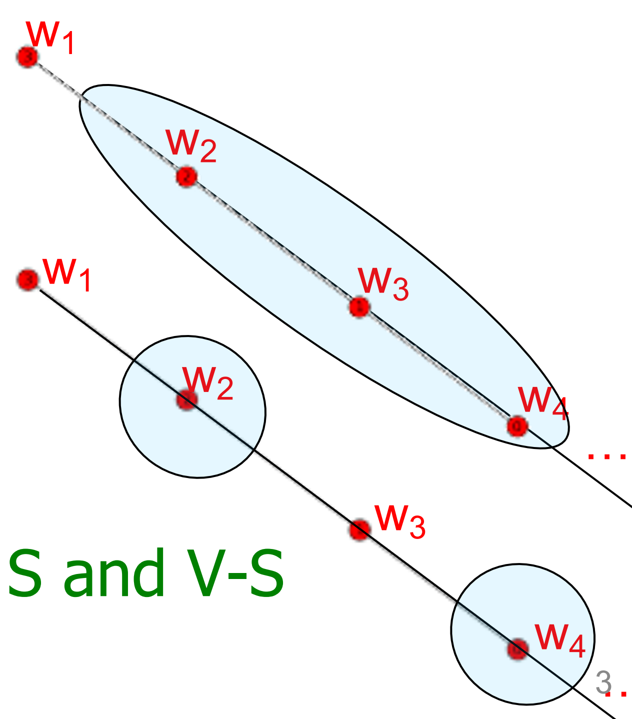
- Ω now is essentially a function defined on a set, not on a vector
  - Set of features with non-zero feature weights w_f
  - E.g. Ω({f_2, f_3, f_4})=2 or Ω({f_2, f_4})=4
  - Detailed values of weights are not what matters to us

Set functions

- Ω now is essentially a function defined on a set
  - Set of features with non-zero feature weights w_f
  - E.g. $\Omega(\{f_2, f_3, f_4\})=2$
- We have a large set V (the universe), Ω is defined for any subset of V, i.e., $\Omega: 2^V \to R$

  (2^V denotes a set of all subsets of V, including V itself, and empty set Ø)
  - V has an element corresponding to each feature, $V=\{f_1, f_2, f_3, f_4,\ldots\}$
- How is Ω defined?
  - We have a graph with V as nodes
  - For an input set S:

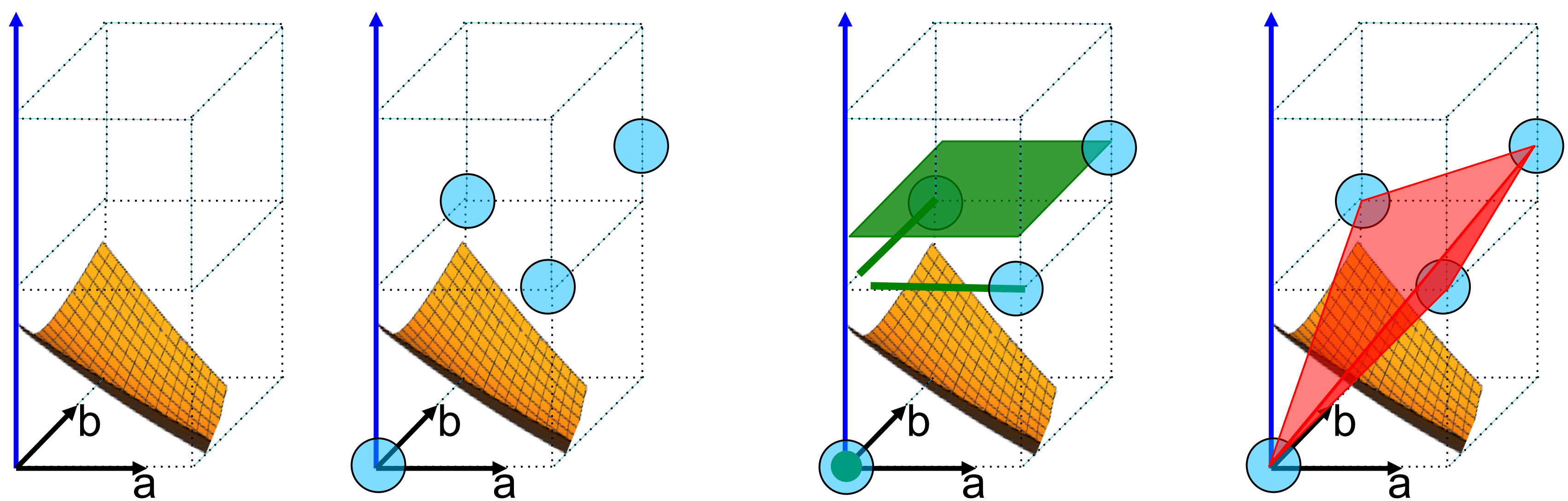
    Ω(S) = number of edges between S and V-S



Big picture

Submodular set functions => convex regularizers!

- Example: two features: x_a and x_b
- Any linear classifier is $y=\text{sign}(w_a x_a + w_b x_b)$
- What are the weights $w=(w_a, w_b)$?
  - To find out, we optimize Risk(w)+Penalty(w)



Risk(w)

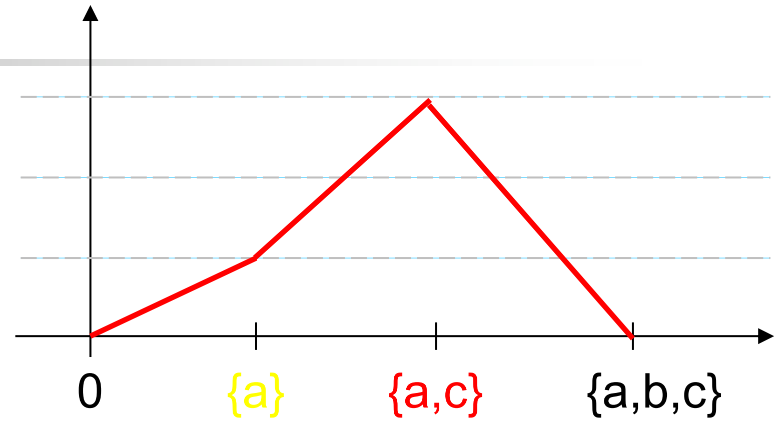
We can't just add penalty on the set (of features) $\Omega(S)$ to risk over vectors (of feature weights).

What we want:
Risk(w)+Penalty(w)
Penalty(w)=$\Omega([w])$
tough to solve!

What we can:
Risk(w)+Penalty(w)
Penalty(w)=$\Omega^{-}(w)$
both convex, so often easy to solve

Set functions

- Set function over universe V, ie. $\Omega: 2^V \to R$
  - E.g. $\Omega_{GC}(S)$ = **graph cut capacity** = (S, V-S) = sum of weights of edges between S and V-S (or just number of edges)
  - Assumptions:
    - weights are positive
    - no edge = zero weight edge

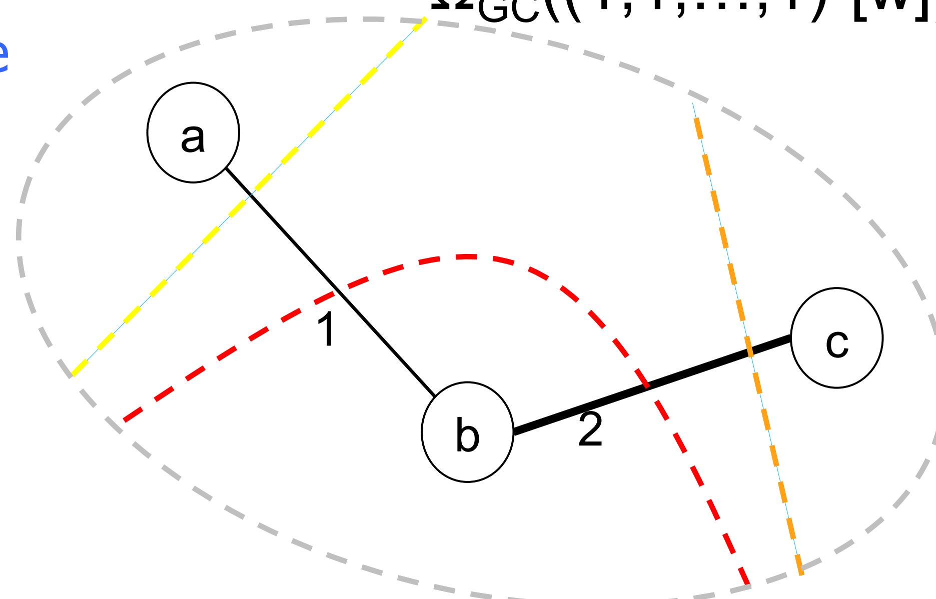


Ω_{GC} is symmetric:

$\Omega_{GC}(S)=\Omega_{GC}(V-S)$

$\Omega_{GC}([w])$

$=\Omega_{GC}((1,1,\ldots,1)-[w])$



E.g. V={a,b,c} with this graph:

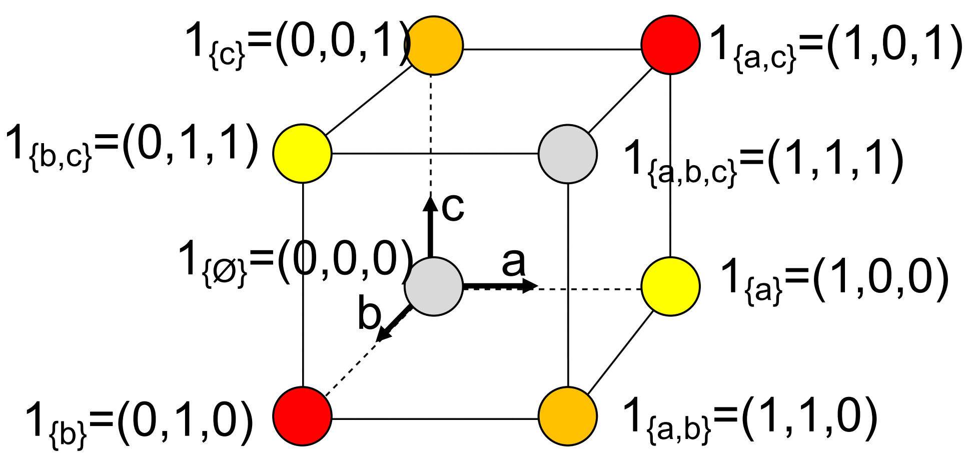
- $\Omega_{GC}(\{\varnothing\}) = \Omega_{GC}(\{a,b,c\})=0$
- $\Omega_{GC}(\{a\}) = \Omega_{GC}(\{b,c\}) =1$
- $\Omega_{GC}(\{c\}) = \Omega_{GC}(\{a,b\}) =2$
- $\Omega_{GC}(\{b\}) = \Omega_{GC}(\{a,c\}) =3$

Set functions

- **We can use binary cube to visualize a set function**

E.g. V={a,b,c} with this graph:

- $\Omega_{GC}(\{\emptyset\}) = \Omega_{GC}(\{a,b,c\})=0$
- $\Omega_{GC}(\{a\}) = \Omega_{GC}(\{b,c\}) =1$
- $\Omega_{GC}(\{c\}) = \Omega_{GC}(\{a,b\}) =2$
- $\Omega_{GC}(\{b\}) = \Omega_{GC}(\{a,c\}) =3$



Each dimension corresponds to an element in the universe V

Notation:

vector to set:

[w]=support of w

vector of 0's and 1's

can be interpreted as a set

set to vector:

1_S = vector with 1's for coordinates corresponding to members of S, and 0's elsewhere

Often we simplify notation and just use set=vector

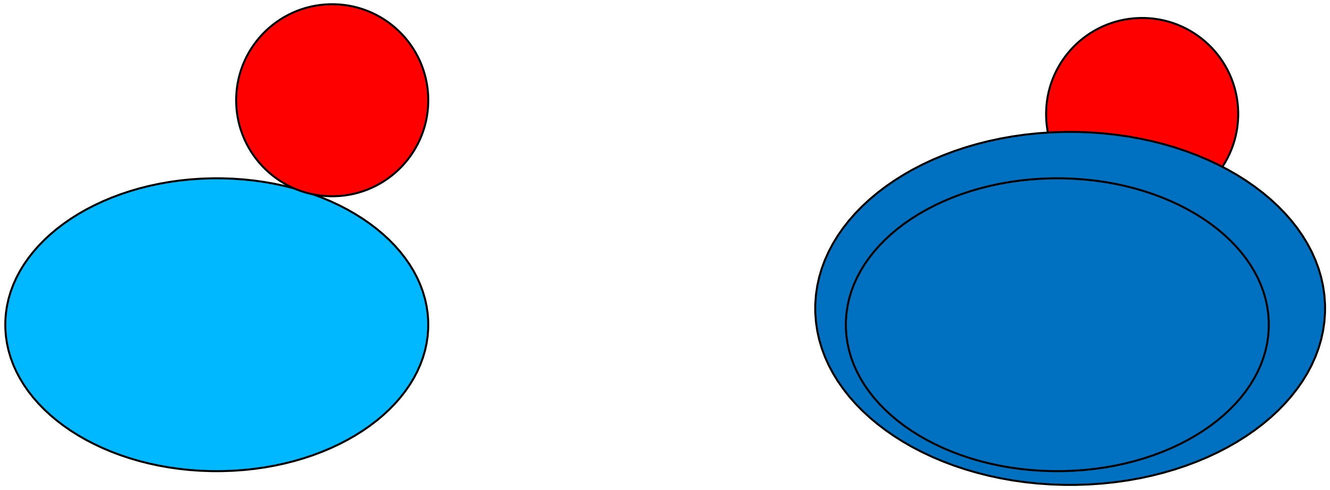
$\Omega(S)=\Omega(1_S)=\Omega([w])$

Submodularity

- Set function over universe V, ie. $\Omega: 2^V \rightarrow R$
- Ω_{GC} is an example of family of set functions called **submodular set functions**

Diminishing returns: 10th slice of cake is not as good as the first one!

f(red+blue) - f(blue) >= f(red+navy) - f(navy)

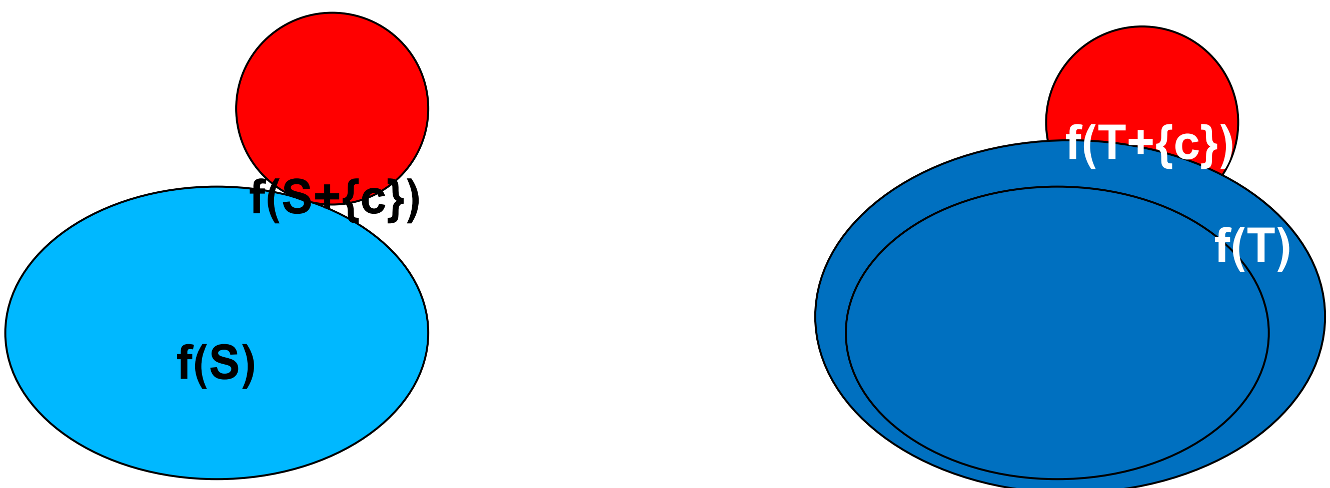


Submodularity

- Set function over universe V, ie. $\Omega: 2^V \rightarrow R$
- Ω_{GC} is an example of family of set functions called **submodular set functions**

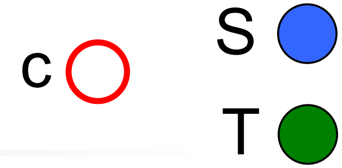
$$\forall S, T, \{c\} \subseteq V,\ S \subseteq T,$$
$$\Omega(S \cup \{c\}) - \Omega(S) \geq \Omega(T \cup \{c\}) - \Omega(T)$$

f(red+blue) - f(blue) >= f(red+navy) - f(navy)

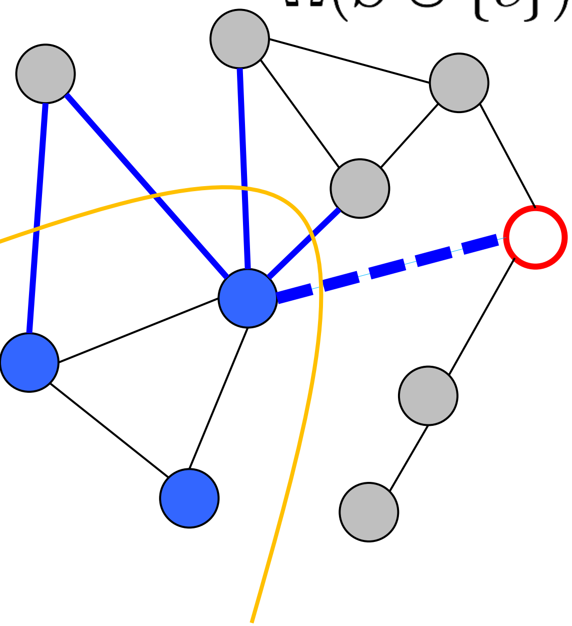


- Adding the same element {c} to a set T increases Ω less then adding it to a subset S of T

Is Graph cut submodular?



$$\Omega(S \cup \{c\}) - \Omega(S) \geq \Omega(T \cup \{c\}) - \Omega(T)$$



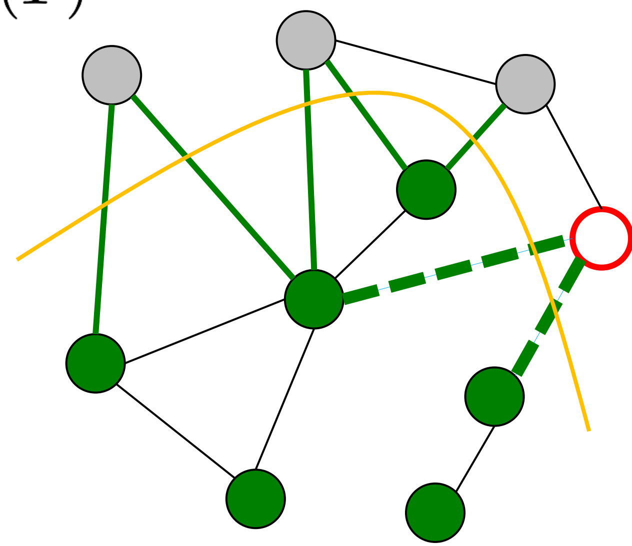
Ω(S)=5

number of edges crossing the cut from S to V-S

To check if graph cut is submodular, we need:

- Set T
- Set S, a subset of T
- A vertex c
  (not a member of T, otherwise we have equality)

- Then, we observe what happens to Ω(S) and Ω(T) when we add {c}



Ω(T)=(T,V-T)=6

number of edges crossing the cut from T to V-T

Is Graph cut submodular?

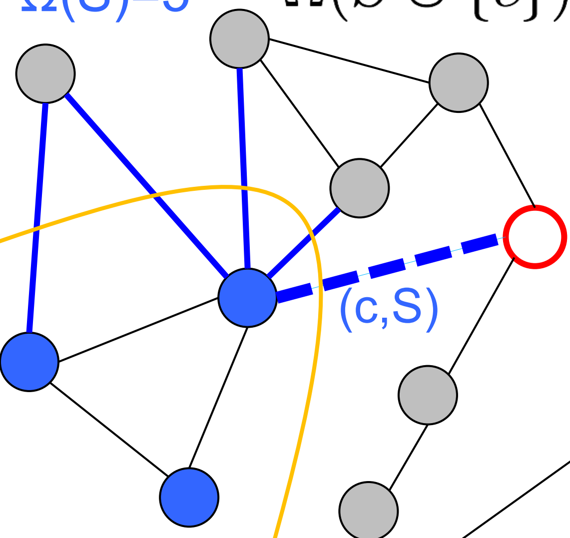
c S T

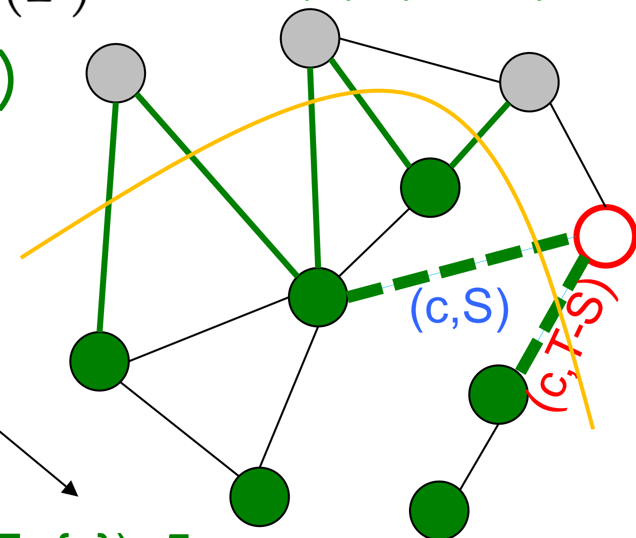
$\Omega(S)=5$

$$\Omega(S \cup \{c\}) - \Omega(S) \geq \Omega(T \cup \{c\}) - \Omega(T)$$

$\Omega(T)=(T,V-T)=6$

… +(c,T-S) ≥ … −(c,T-S)

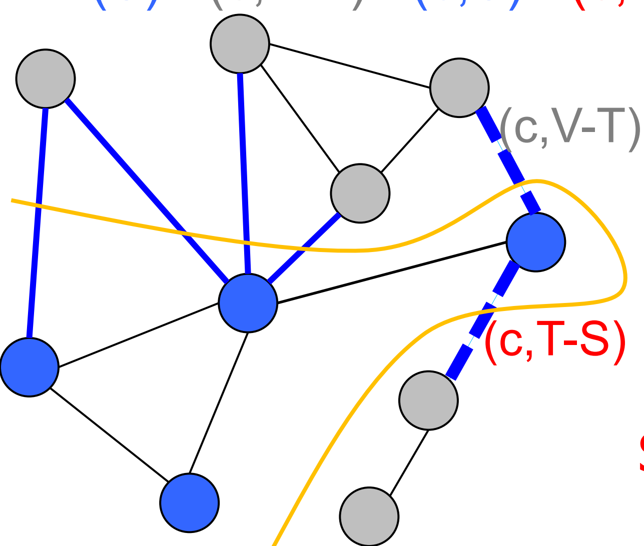




what happens to
$\Omega(S)$ and $\Omega(T)$
when we add {c}

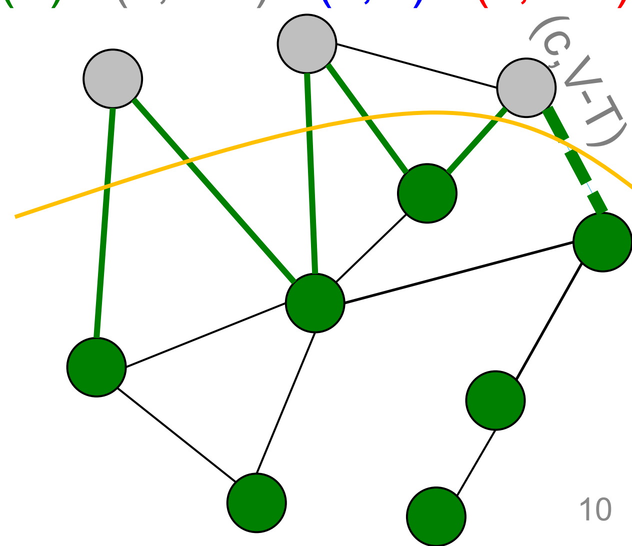
$\Omega(S+\{c\})=6$
$=\Omega(S)$ +(c,V-T) −(c,S) +(c,T-S)

$\Omega(T+\{c\})=5$
$=\Omega(T)$ +(c,V-T) −(c,S) −(c,T-S)



If vertex c is connected
to vertices in (T-S)
with nonzero
sum of weights (c,T-S),
we have "≥" inequality

So: graph cut capacity is a
submodular set function

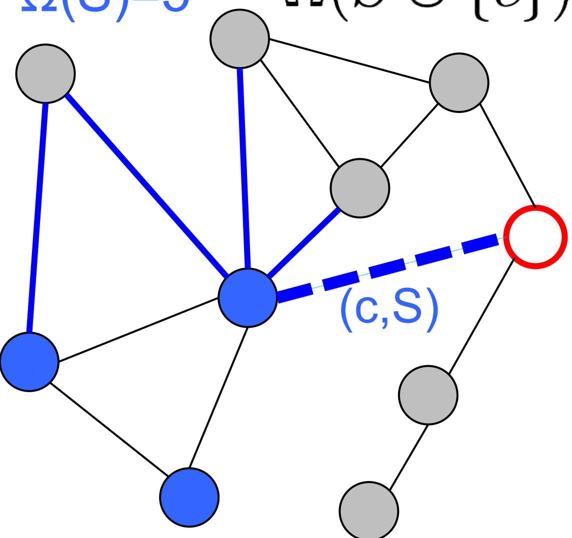


Is Graph cut submodular?

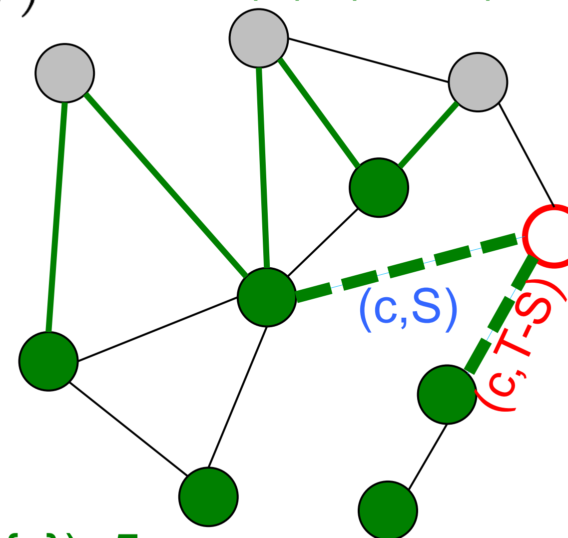
c S T

$$\Omega(S \cup \{c\}) - \Omega(S) \geq \Omega(T \cup \{c\}) - \Omega(T)$$

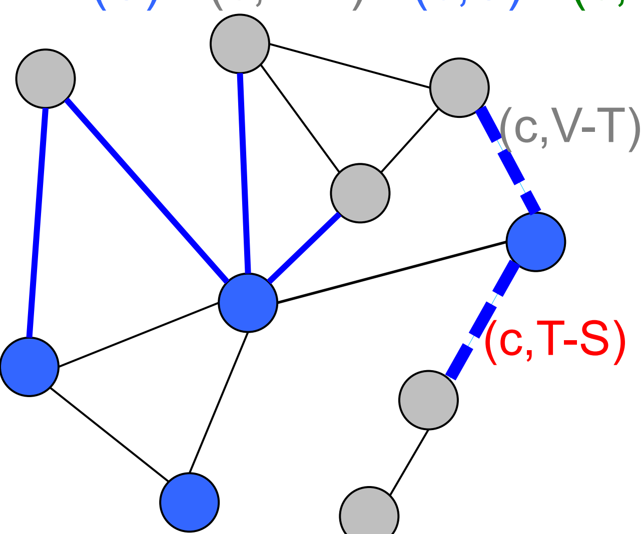
Ω(S)=5



Ω(T)=(T,V-T)=6

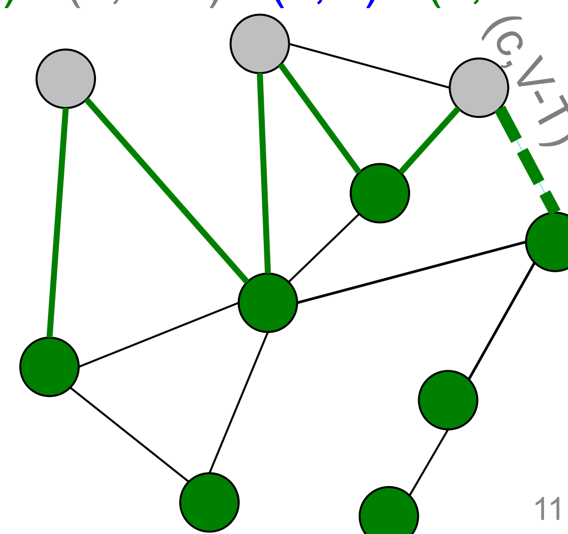


Ω(S+{c})=6
=Ω(S) +(c,V-T) –(c,S) +(c,T-S)

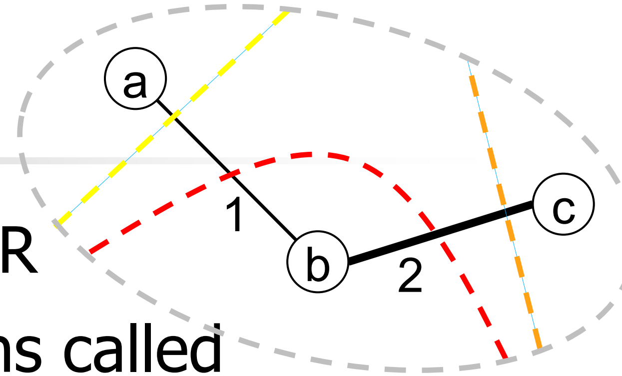


We have equality
when (c,T-S)=0, i.e.,
when there are no edges
between vertex c
and vertices in (T-S)

Ω(T+{c})=5
=Ω(T) +(c,V-T) –(c,S) –(c,T-S)



Submodularity

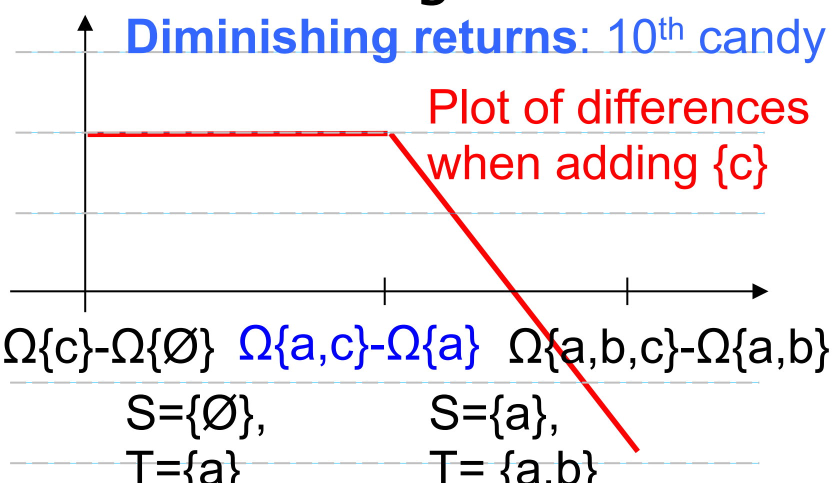


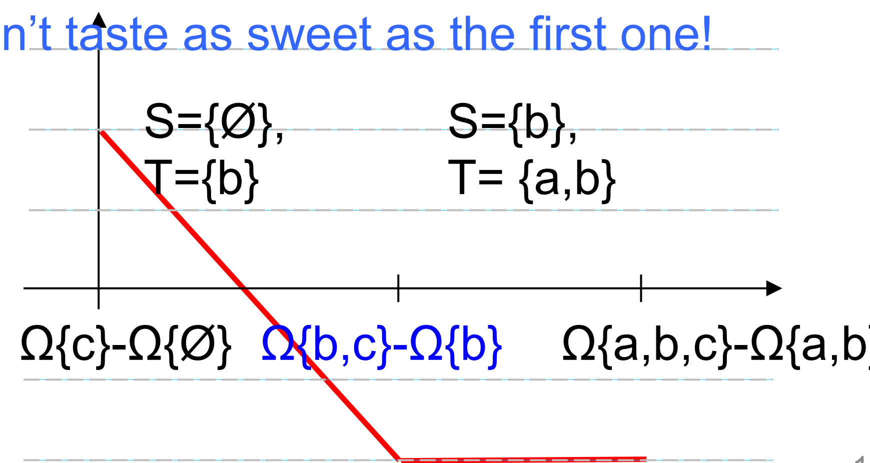
- Set function over universe V, ie. $\Omega: 2^V \to R$
- Ω_{GC} is an example of family of set functions called **submodular set functions**
  - Set function $\Omega: 2^V \to R$ is **submodular** iff:

$$\forall S, T, \{c\} \subseteq V,\ \ S \subseteq T,$$
$$\Omega(S \cup \{c\}) - \Omega(S) \geq \Omega(T \cup \{c\}) - \Omega(T)$$

  - Adding the same element {c} to a set T increases Ω less then adding it to a subset S of T

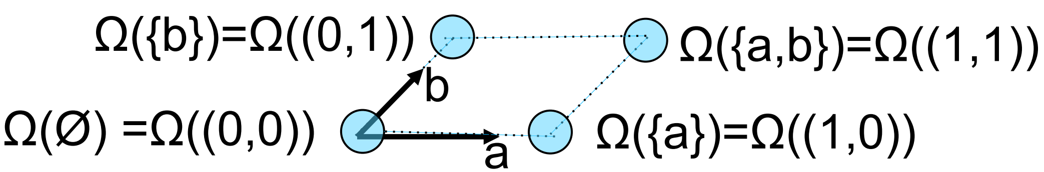
Diminishing returns: 10^{th} candy doesn't taste as sweet as the first one!

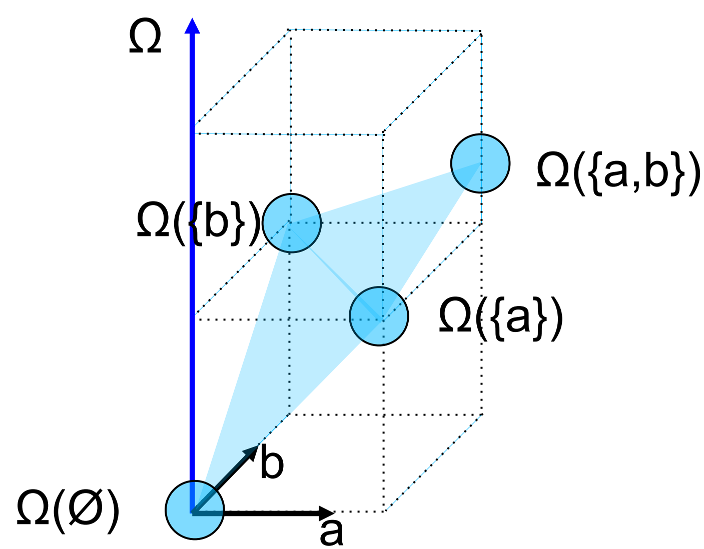




Submodular set functions

- Submodular set functions are a bit similar to concave functions - they're "bending down"
  - we can plot the function on {a,b} on a unit cube
  - value Ω({a,b}) is lower than Ω({a})+Ω({b})
    - S=Ø, T={a}, c=b
    - Assume w.l.o.g. Ω(Ø)=0





submodular Ω

$$\forall S, T, \{c\} \subseteq V,\ \ S \subseteq T,$$
$$\Omega(S \cup \{c\}) - \Omega(S) \geq \Omega(T \cup \{c\}) - \Omega(T)$$

Alternative definitions

- **Modular** set functions: $\Omega(\{a,b,\ldots,z\})=\Omega_a+\Omega_b+\ldots+\Omega_z$

$$\forall A,B \subseteq V \quad \Omega(A\cup B)+\Omega(A\cap B)=\Omega(A)+\Omega(B)$$

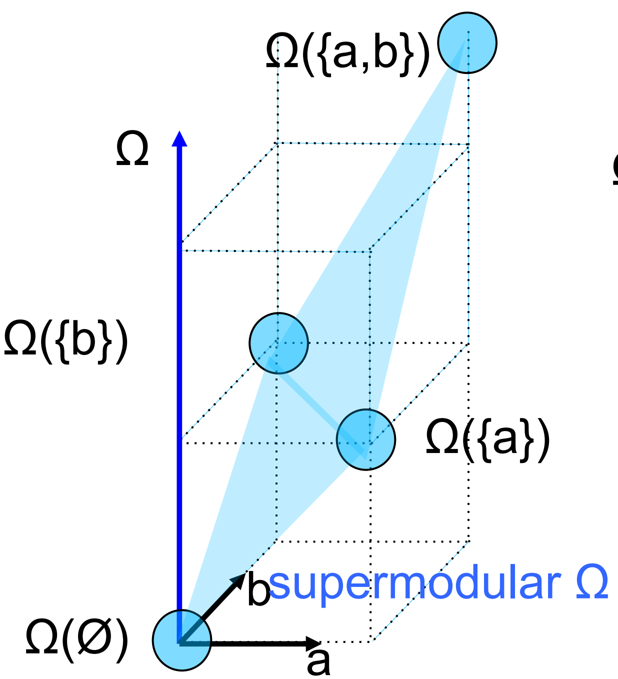
- **Sub**modular set functions: $\Omega(\{a,b,\ldots,z\}) \leq \Omega_a+\Omega_b+\ldots+\Omega_z$

$$\Omega(A\cup B)+\Omega(A\cap B)\leq\Omega(A)+\Omega(B)$$

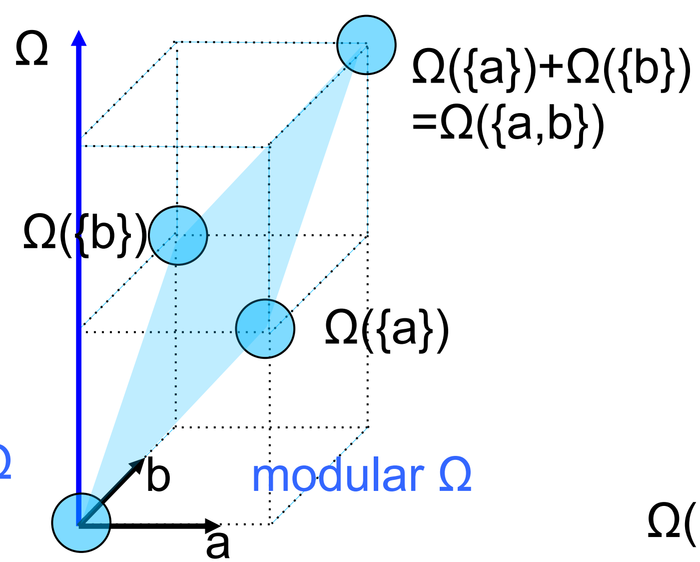
- **Super**modular set functions: $\Omega(\{a,b,\ldots,z\}) \geq \Omega_a+\Omega_b+\ldots+\Omega_z$

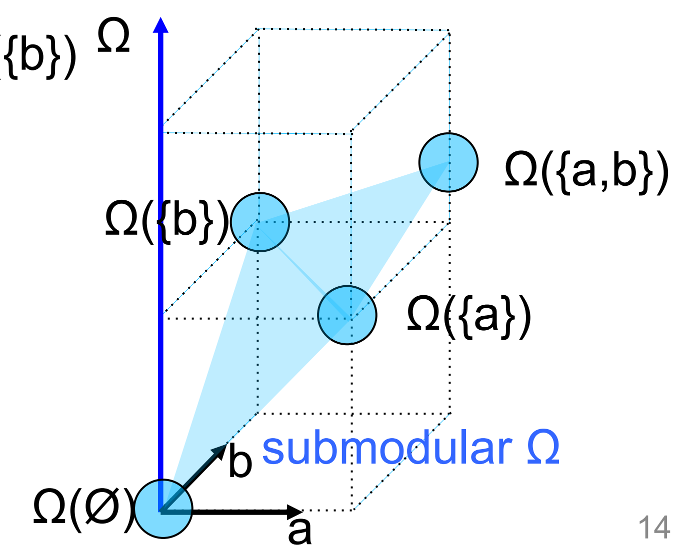
$$\Omega(A\cup B)+\Omega(A\cap B)\geq\Omega(A)+\Omega(B)$$

- If Ω is modular, it is both submodular and supermodular



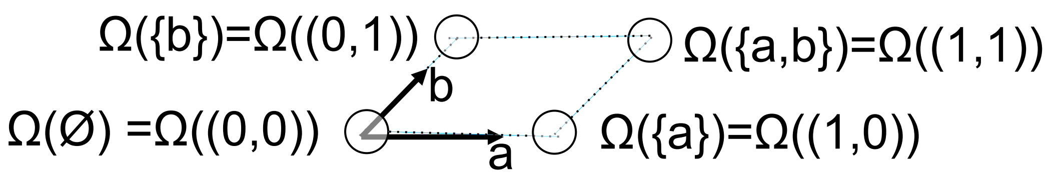
- If Ω is submodular, -Ω is supermodular



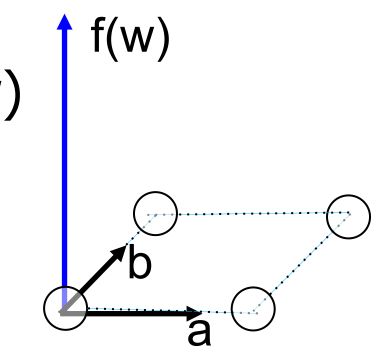


Submodular set functions

- Set functions: defined vectors with coordinates $\{0,1\}^F$, e.g. $\{a,c\}=(1,0,1)$
  - i.e. vectors at the corner of unit cube $\{0,1\}^F$ where $F=|V|$ is number of elements in V



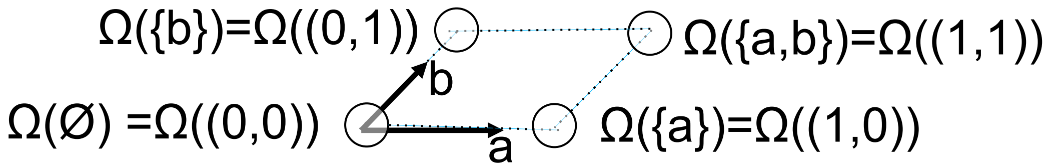
- Given a **set** function Ω can we define a function f(w)
  $f: [0,1]^F \to R$
  (or even better, $R^F \to R$)?



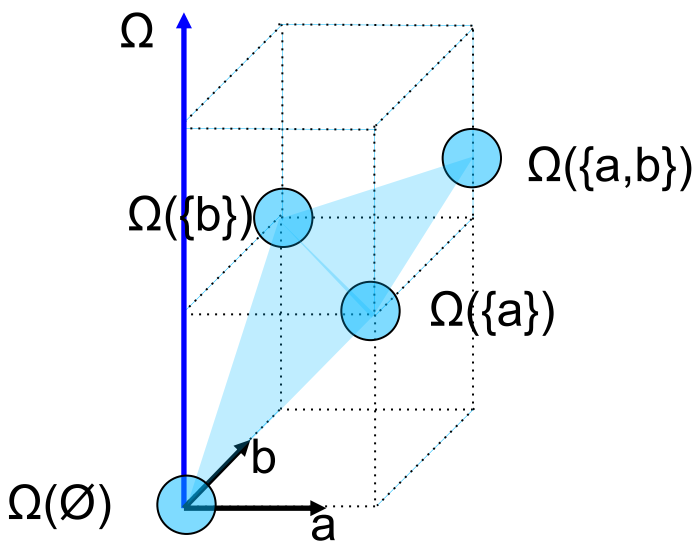
- A function that would be somehow related to Ω
  - We already did!
  - When discussing notation:
  - f(w)=Ω([w]) *f = # of non-zero elements in w*
    - For any vector R^F, we get vector of 0's/1's, which represents a set, and defines value of f(w)
    - What's wrong with using f(w)=Ω([w])?

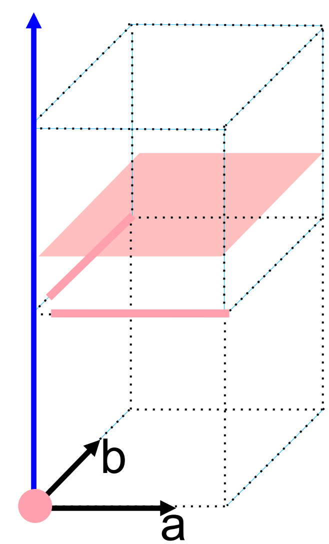
Submodular set functions

- Set functions: defined vectors with coordinates {0,1}, e.g. {a,c}=(1,0,1)
  - i.e. vectors at the corner of unit cube $\{0,1\}^F$ where F=|V| is number of elements in V



- Given a set function Ω can we define a "nicely behaving" function f: $[0,1]^F$ -> R (or even better, R^F -> R)?
- A function that would be somehow related to Ω
  - E.g. the light blue piecewise linear function on [0,1] x [0,1] we've seen before:





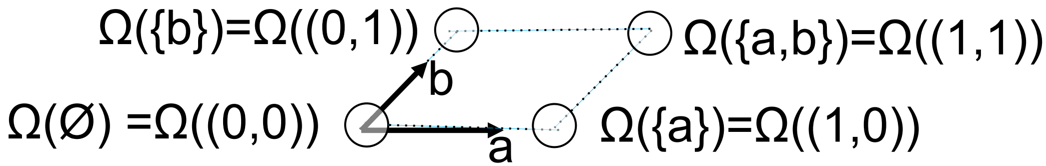
f(w)=Ω([w])

Not continuous!

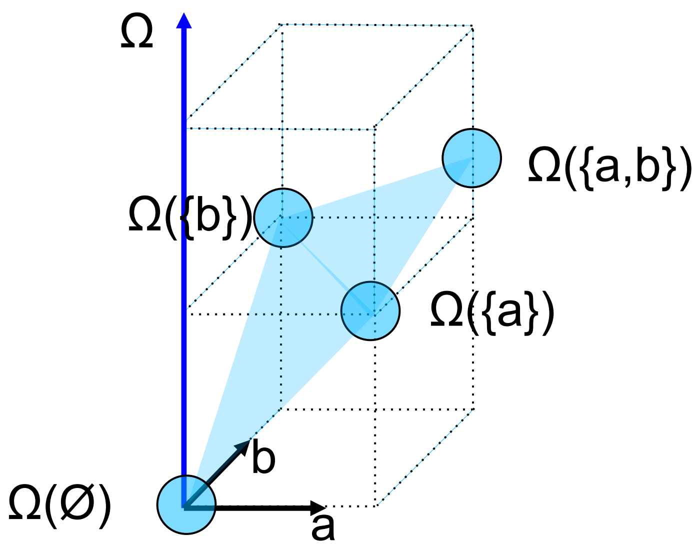
Not convex!
e.g. try (1,0) and (0,1) and points in between

Submodular set functions

- Set functions: defined vectors with coordinates {0,1}, e.g. {a,c}=(1,0,1)
  - i.e. vectors at the corner of unit cube $\{0,1\}^F$ where F=|V| is number of elements in V

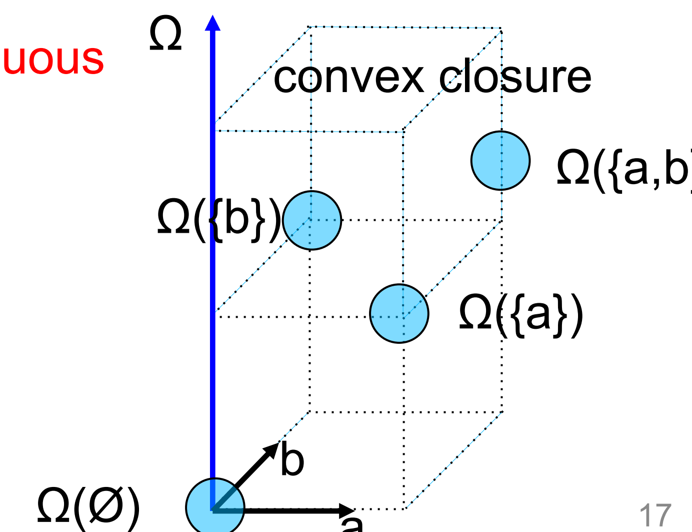


- Given a set function Ω can we define a "nicely behaving" function f: $[0,1]^F$ -> R (or even better, R^F -> R)?
- A function that would be somehow related to Ω
  - E.g. the light blue piecewise linear function on [0,1] x [0,1] we've seen before:



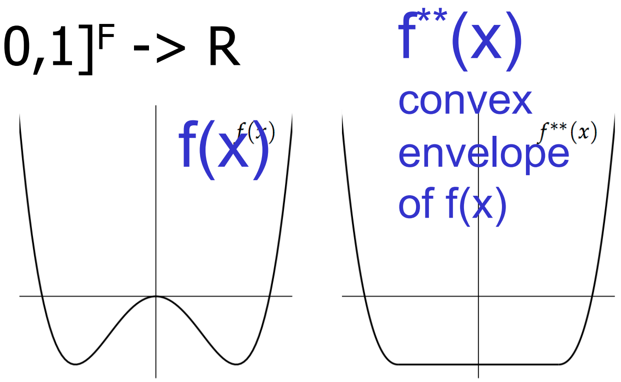
<- this one is continuous

But is it convex?

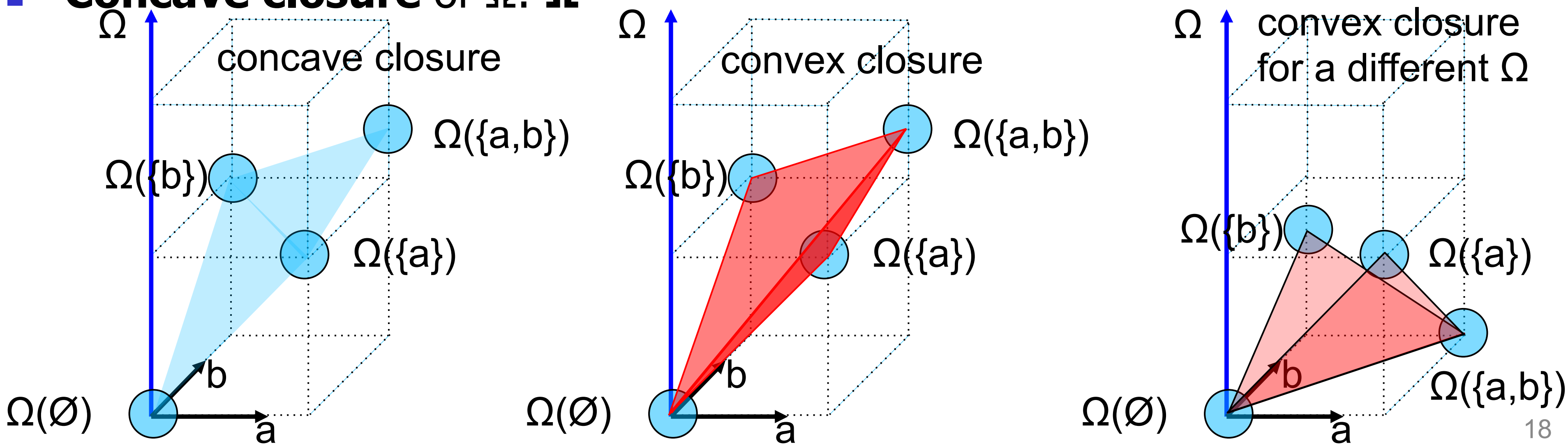


Submodular set functions

- Given a set function Ω we define two functions $f: [0,1]^F \to R$
- **Convex closure** of Ω: $\boldsymbol{\Omega^-}$
  - It is the **convex envelope (**Ω^{**}) of Ω: pointwise highest convex function that bounds Ω from below: for any w in $\{0,1\}^F$: $\Omega^-(w) \leq \Omega(w)$
  - Convex closure of set function is **piecewise linear**: Intuition: Start with any convex function that bounds Ω from below, push every point of it up until you can't without loosing convexity – you will get maxima of hyperplanes

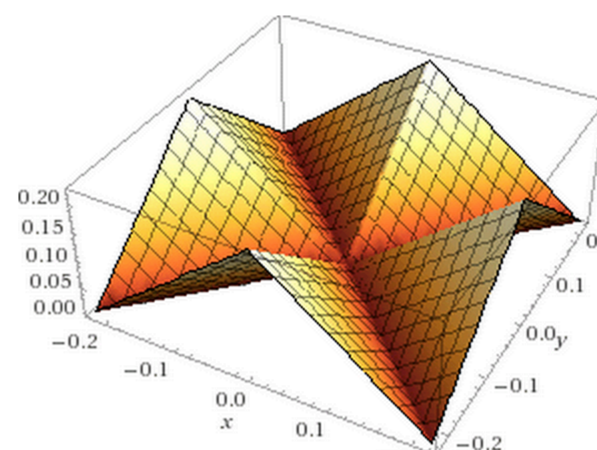


- **Concave closure** of Ω: $\boldsymbol{\Omega^+}$



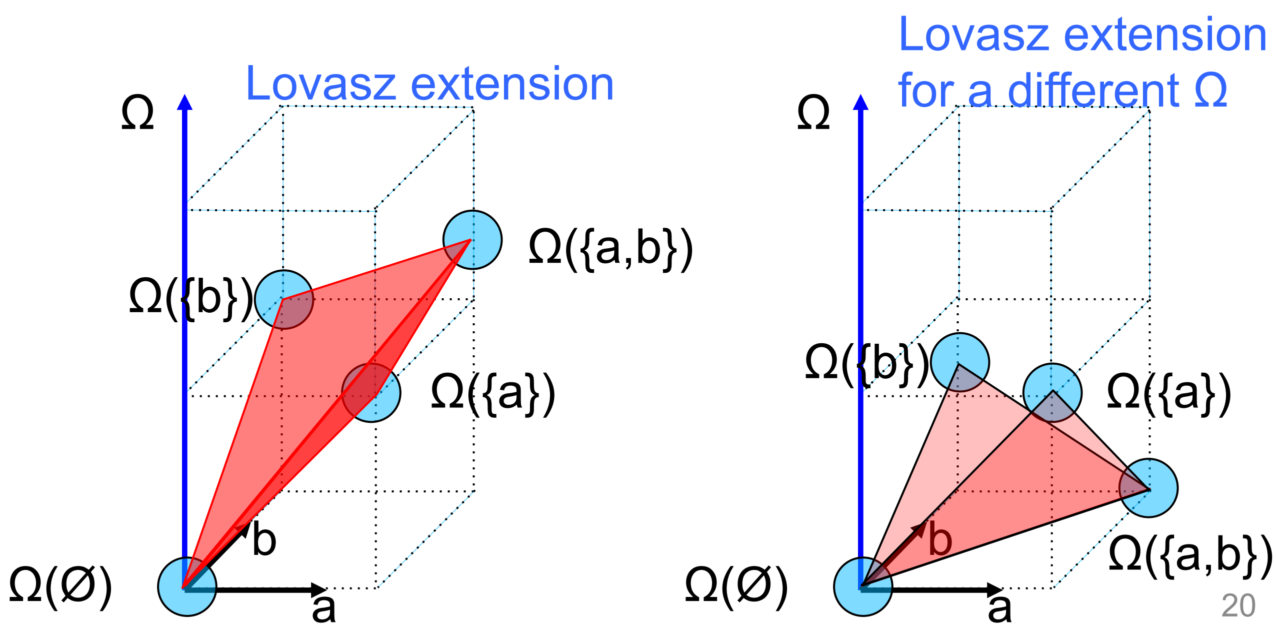
Extensions vs. envelopes

- **Extension of set function:**
- For any set function Ω (submodular or not): $\boldsymbol{\Omega^-([w])=\Omega([w])}$
- **Convex closure Ω^-** is an **extension** of set function **Ω**, that is, has the same values as **Ω** on the corners of unit cube $[0,1]^F$
- This is quite remarkable!
  - It's possible because at each dimension, we only have value of **Ω** at two points, 0 and 1, so we can connect them by straight line (which is convex/concave)
  - Imagine we had a function defined over three not two points in every dimension: $\{-1,0,1\}^F$
  - For most of such functions, convex envelope would not go through values of function at the points $\{-1,0,1\}^F$
  - Conversely, extension would not be convex
  - E.g. here, extension is not convex, convex envelope is f(w)=0 for any w



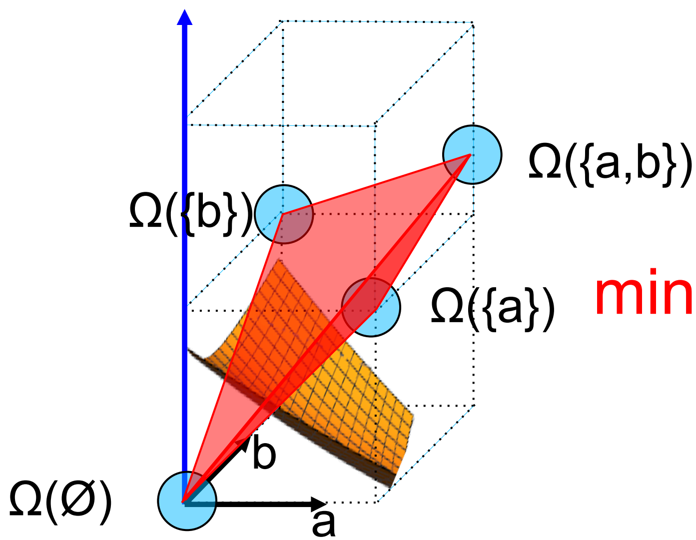
Submodular minimization

- Given a submodular set function $\Omega(S)=\Omega(w)$ where $S=1_w$
- We have a well-defined Lovasz extension Ω^L: $[0,1]^n$ -> R
  - For a binary vector w representing set S: $\Omega^L(w)=\Omega(S)$
  - Ω^L is continuous, convex, piecewise linear
  - $\Omega^L = \Omega^-$ Pointwise highest convex function bounding Ω from below
- $\Omega(S)$ and $\Omega^L(w)$ have the same global minimum
  - And minimum is attained at a corner: at some **w** in $\{0,1\}^F$



Submodular minimization

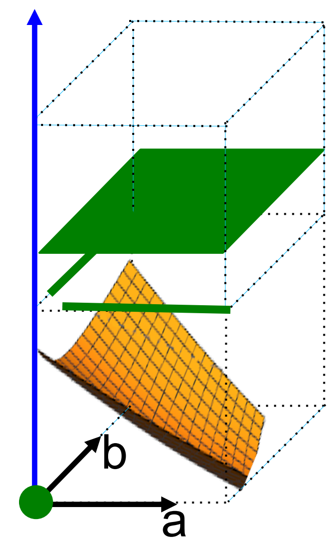
- Given a set function $\Omega(S)=\Omega(w)$ where $S=1_w$
- We have a well-defined Lovasz extension Ω^L: $[0,1]^n$ -> R



$\Omega^L(w) \neq \Omega([w])$

$\min_w \Omega^L(w) = \min_w \Omega([w])$

however:



$\min_w f(w)+\Omega^L(w) \neq \min_w f(w)+\Omega([w])$

- Consequence: $\mathrm{argmin}_w\, R_S(w)+\Omega^L(w) \neq \mathrm{argmin}_w\, R_S(w)+\Omega([w])$

Overall approach

- For any submodular set function $\Omega(1_S)$: $\{0,1\}^F \to R$
  we can construct its extension $\Omega^L(w)$: $[0,1]^F \to R$
    - And then define $\Omega^L(w)$: $R^F \to R$
- $\text{argmin}_w R_S(w)+\Omega^L(w) \neq \text{argmin}_w R_S(w)+\Omega([w])$
  - But still, $\Omega^L(w)$ captures some aspects of Ω
  - E.g. Penalty=L_1 norm vs. Penalty=# of features used
- Unlike $\Omega([w])$, the extension $\Omega^L(w)$ is continuous, convex (though typically non-differentiable)
- We have tool for dealing with non-differentiable, convex terms:
  - Proximal gradient descent, we need: $\text{argmin}_w \Omega^L(w)+b||w - v||^2$
- So, we can solve problems of the form:
  - **Differentiable risk + convex extension of a submodular set function Ω**

Big picture

- We can solve problems of the form:
  - **Differentiable risk**
    + convex extension of a submodular set function Ω

- What are some interesting submodular set functions?
  - What are their extensions?
  - Can we solve proximal operator for them?