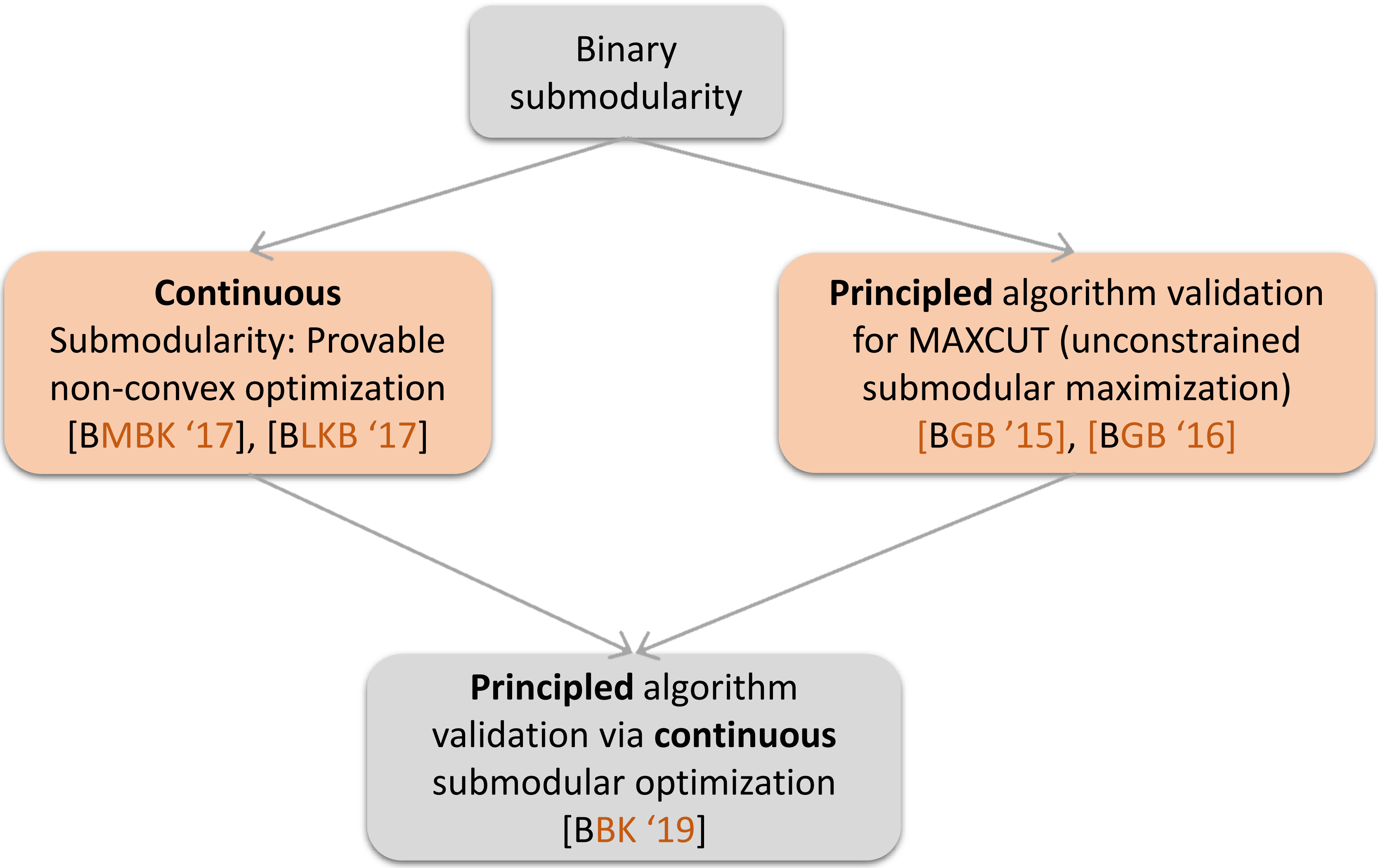
ETH zürich

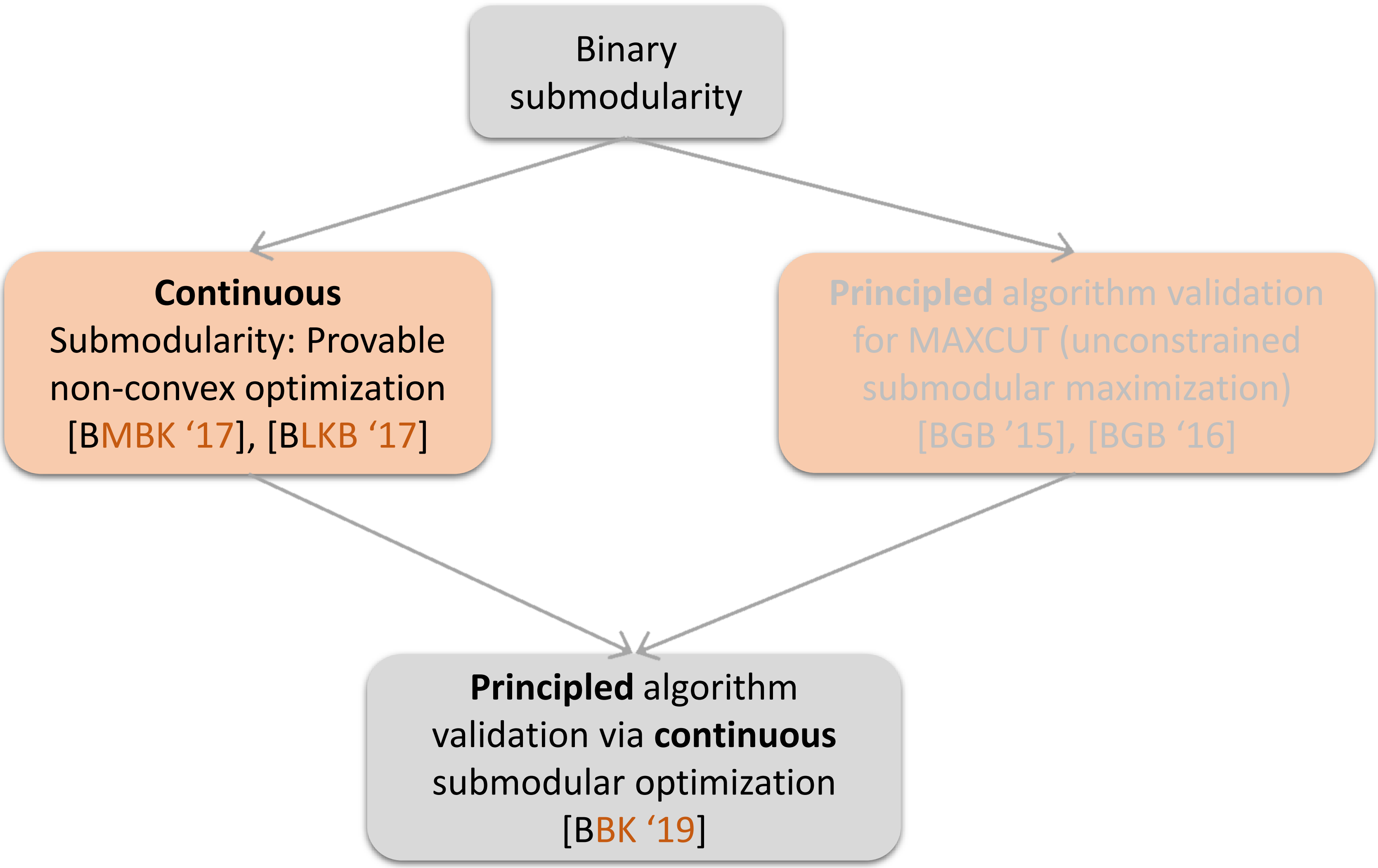
Provable Non-Convex Optimization and Algorithm Validation via Submodularity

Yatao Bian

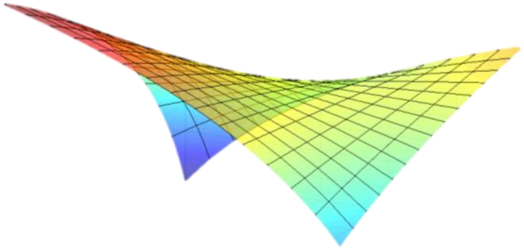
ETH Zürich

November 20, 2019


Binary submodularity
Continuous Submodularity: Provable non-convex optimization [BMBK ‘17], [BLKB ‘17]
Principled algorithm validation for MAXCUT (unconstrained submodular maximization) [BGB ’15], [BGB ‘16]
Principled algorithm validation via **continuous** submodular optimization [BBK ‘19]


Binary submodularity
Continuous Submodularity: Provable non-convex optimization [BMBK '17], [BLKB '17]
Principled algorithm validation for MAXCUT (unconstrained submodular maximization) [BGB '15], [BGB '16]
Principled algorithm validation via **continuous** submodular optimization [BBK '19]

Why do we need continuous submodularity?



Motivations and applications

Motivation 1: Prior knowledge for modeling

Continuous DR-submodularity captures
"Diminishing Returns (DR)" phenomenon

$f($   $)$

happiness gained by having
some quantity of (water, coke)

$$\boldsymbol{\delta} = \begin{bmatrix} 50\ ml \text{ water} \\ 50\ ml \text{ coke} \end{bmatrix}$$

$$f\left(\boldsymbol{\delta} + \begin{bmatrix} 1\ ml \\ 1\ ml \end{bmatrix}\right) - f\left(\begin{bmatrix} 1\ ml \\ 1\ ml \end{bmatrix}\right)$$

$$\geq f\left(\boldsymbol{\delta} + \begin{bmatrix} 100\ ml \\ 100\ ml \end{bmatrix}\right) - f\left(\begin{bmatrix} 100\ ml \\ 100\ ml \end{bmatrix}\right)$$

marginal gain of happiness by
having $\boldsymbol{\delta}$ more (water, coke)
based on a large context

To model:
- preference
- influence
- satisfaction
- revenue
…

Motivation 2: A non-convex structure with provable optimization

Quadratic Program (QP):

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{H}\mathbf{x} + \mathbf{h}^{\mathrm{T}}\mathbf{x} + c,\ \mathbf{H} \text{ is symmetric}$$

$$\nabla^2 f = \mathbf{H},\ \mathbf{H} = \begin{bmatrix} -1 & -2 \\ -2 & -1 \end{bmatrix},\ \text{eigenvalues: } \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

Non-convex/non-concave → Continuous submodular 😃

It arises in:

Lovasz/Multilinear extensions of submodular set functions [Lovasz '83][Calinescu et al. '07]

DPP MAP inference [Gillenwater et al. '12] [BLKB '17]

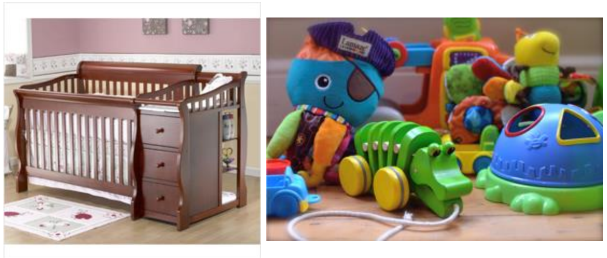
Social network mining [BMBK '17]

Risk-sensitive submodular optimization [Wilder '17]

Robust budget allocation [Staib et al. '17]

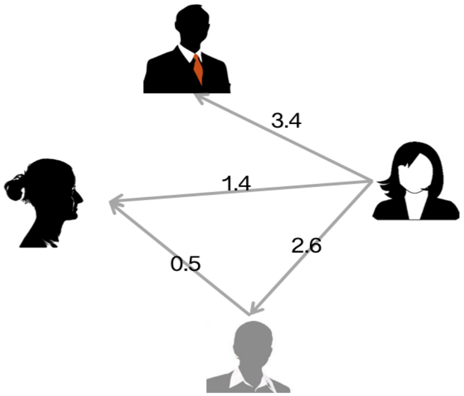
Mean field inference for the posterior agreement (PA) distribution [BBK '19]

Product recommendation



Amazon baby registries. Left: furniture, right: toys

Revenue maximization



Budget allocation

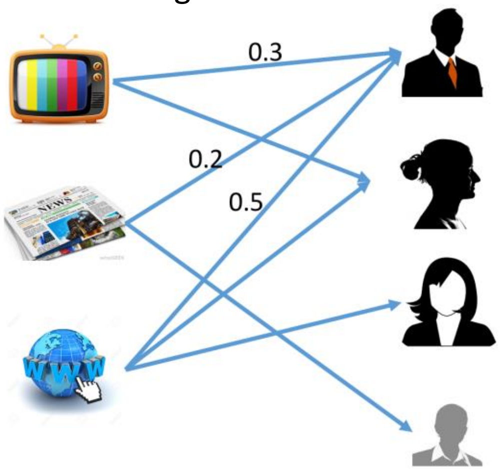
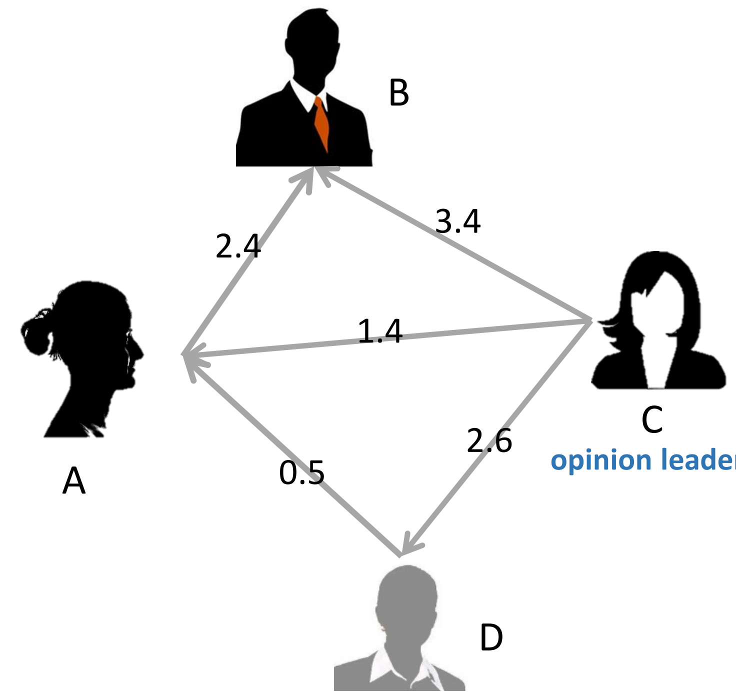


Image summarization

Revenue maximization with continuous assignments

[Hartline et al. ’08, BMBK ‘17]

Task: advertise an innovation/product based on a social connection graph
→ max. expected revenue



Given: connection graph

- Nodes: all users (all people on FB)
- Edges: influence strength between users

Viral marketing: give some users a certain amount of free products, to trigger further adoptions

$\mathbf{x} \in \mathbb{R}_+^n$: free trial time for n users

How to model expected revenue: $f(\mathbf{x})$?

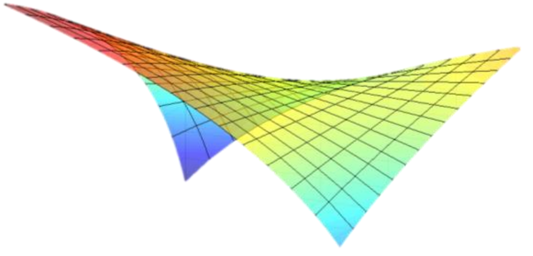
Revenue $f(\mathbf{x})$ satisfies DR property:

giving  more trial time

will hurt influence of 

How to characterize continuous submodularity?

Definitions and characterizations



Three orders of characterizations [BMBK '17]

coordinate-wise less equal

antitone mapping: $\mathbf{x} \lesssim \mathbf{y}$ implies $\nabla f(\mathbf{x}) \gtrsim \nabla f(\mathbf{y})$

	Continuous submodular f	Convex g, $\lambda \in [0,1]$
0^{th} order	$f(\mathbf{x}) + f(\mathbf{y}) \geq f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y})$	$\lambda g(\mathbf{x}) + (1-\lambda) g(\mathbf{y}) \geq g(\lambda \mathbf{x} + (1-\lambda)\mathbf{y})$
1^{st} order	$\nabla f(\cdot)$: weak antitone mapping	$g(\mathbf{y}) \geq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$
2^{nd} order	$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \leq 0, \forall i \neq j$	$\nabla^2 g(\mathbf{x}) \succcurlyeq 0$ (PSD)

not care	≤ 0	≤ 0	≤ 0
≤ 0	not care	≤ 0	≤ 0
≤ 0	≤ 0	not care	≤ 0
≤ 0	≤ 0	≤ 0	not care

Hessian

$\vee$: coordinate-wise max.

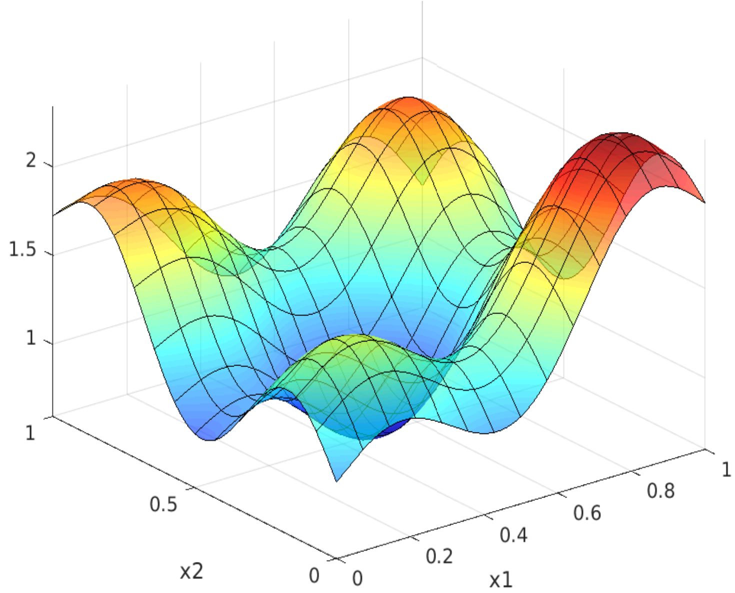
$\wedge$: coordinate-wise min.

$\mathbf{x}$	$\mathbf{y}$	$\mathbf{x} \vee \mathbf{y}$	$\mathbf{x} \wedge \mathbf{y}$
2	1	2	1
0	2	2	0
4	3	4	3

Continuous submodularity: Repulsion among different dimensions

does not say anything about a single coordinate!

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \leq 0, \forall i \neq j$$



Arbitrary behavior along a single coordinate

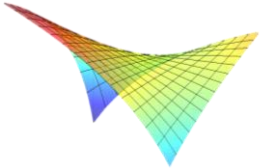
not care	≤ 0	≤ 0	≤ 0
≤ 0	not care	≤ 0	≤ 0
≤ 0	≤ 0	not care	≤ 0
≤ 0	≤ 0	≤ 0	not care

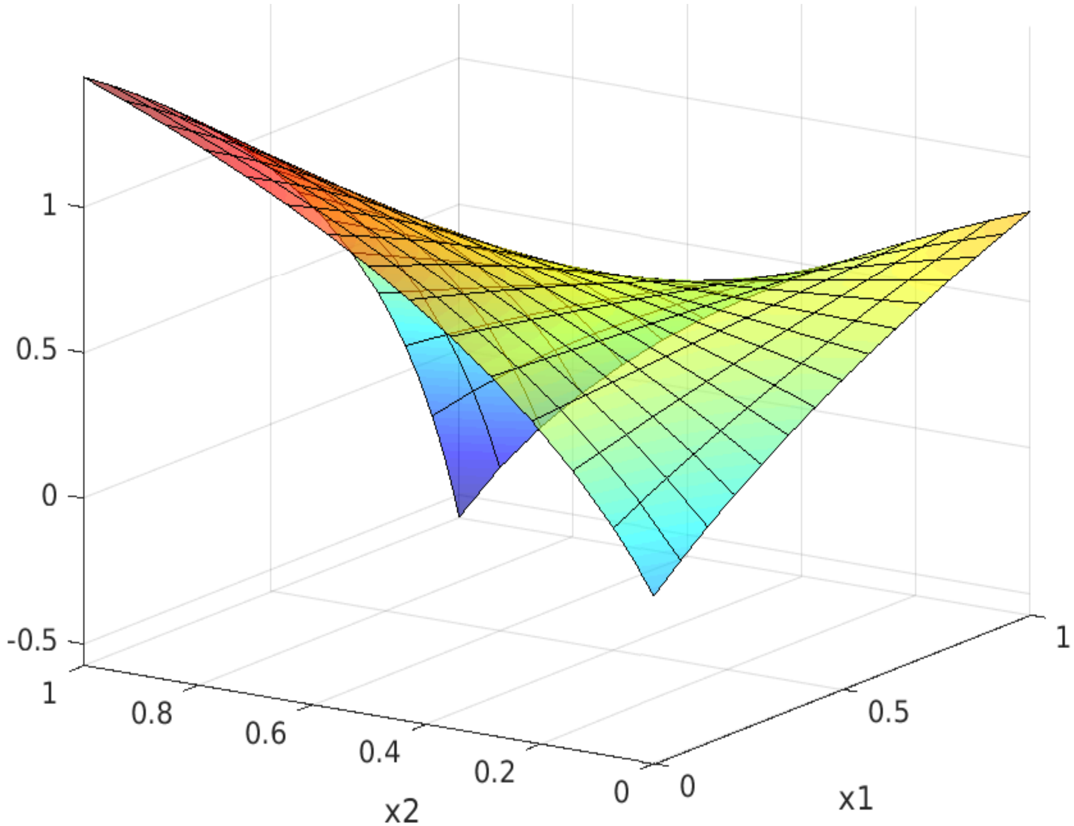
Hessian

Often, objectives have some structure along a single coordinate

Submodularity + Concavity along any single coordinate

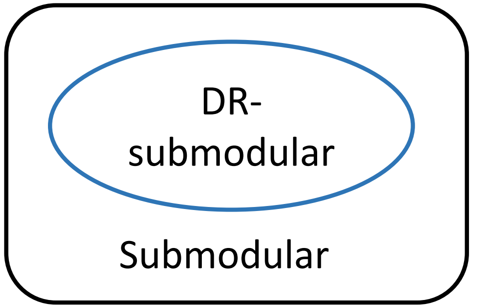
= Continuous DR-submodularity





A 2-D Softmax extension [Gillenwater et al. '12]

Two classes of continuous submodular functions [BMBK '17]

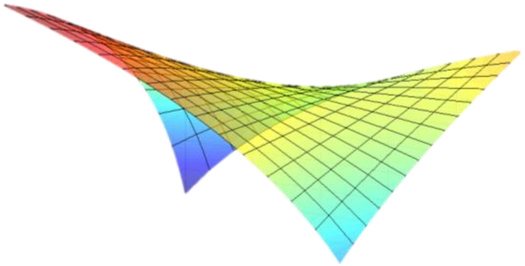


	Continuous Submodular	Continuous DR-Submodular
0th order	$f(\mathbf{x}) + f(\mathbf{y}) \geq f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y})$	$d(\mathbf{x}) + d(\mathbf{y}) \geq d(\mathbf{x} \vee \mathbf{y}) + d(\mathbf{x} \wedge \mathbf{y})$ & coordinate-wise concave
1st order	$\nabla f(\cdot)$: weak antitone mapping	$\nabla d(\cdot)$: antitone mapping
2nd order	$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \leq 0, \forall i \neq j$	$\frac{\partial^2 d(\mathbf{x})}{\partial x_i \partial x_j} \leq 0, \forall i, j$

not care	≤ 0	≤ 0	≤ 0
≤ 0	not care	≤ 0	≤ 0
≤ 0	≤ 0	not care	≤ 0
≤ 0	≤ 0	≤ 0	not care

≤ 0	≤ 0	≤ 0	≤ 0
≤ 0	≤ 0	≤ 0	≤ 0
≤ 0	≤ 0	≤ 0	≤ 0
≤ 0	≤ 0	≤ 0	≤ 0

How to maximize continuous DR-submodular functions?



Provable algorithms

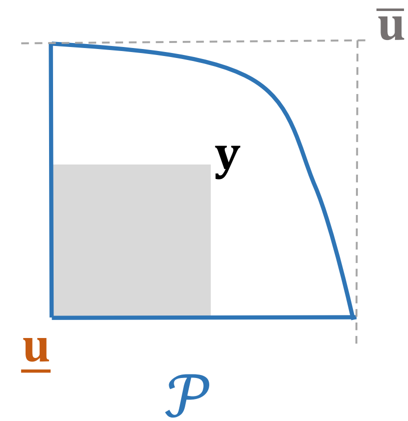
DR-submodular maximization: Setup & hardness

$$\max_{\mathbf{x}\in\mathcal{P}} f(\mathbf{x})$$

$\mathcal{P}$ is convex
& down-closed:

A convex set with a lower bound $\underline{\mathbf{u}}$, s.t.

$\forall \mathbf{y} \in \mathcal{P}$, the hyperrectangle $[\underline{\mathbf{u}}, \mathbf{y}] \subseteq \mathcal{P}$



Hardness & Inapproximability: The above problem is NP-hard. When $\mathcal{P}$ is a unit hypercube ($\mathcal{P}$=$[0,1]^n$), there is no poly. time $(1/2+\varepsilon)$−approximation algorithm for any $\varepsilon > 0$ unless RP=NP.

½-approximation: finding a solution $\mathbf{x}$ s.t. $f(\mathbf{x}) \geq \frac{1}{2} f(\mathbf{x}^*)$

A summary of our theoretical results

Mathematical characterizations of submodularity over integer & continuous domains [BMBK '17]	0^{th} order, 1^{st} order, 2^{nd} order, antitone gradient etc
$\mathbf{a} \lesssim \mathbf{b} \rightarrow f(\mathbf{a}) \leq f(\mathbf{b})$ Monotone DR-submodular max. with down-closed convex constraints [BMBK '17]	- Inapproximability: $1 - 1/e$ **- Optimal** algorithm: A Frank-Wolfe Variant
Non-monotone DR-submodular max. with box constraints [BBK '19]	- Inapproximability: $1/2$ **- Optimal** algorithm: DR-DoubleGreedy
Non-monotone DR-submodular max. with down-closed convex constraints [BLKB '17]	- Inapproximability: Open problem - Best algorithm so far: Shrunken Frank-Wolfe, $1/e$ guarantee

Non-monotone DR-submodular max. with down-closed convex constraints [BLKB '17]

- Inapproximability: Open problem
- Best algorithm so far: Shrunken Frank-Wolfe, $1/e$ guarantee

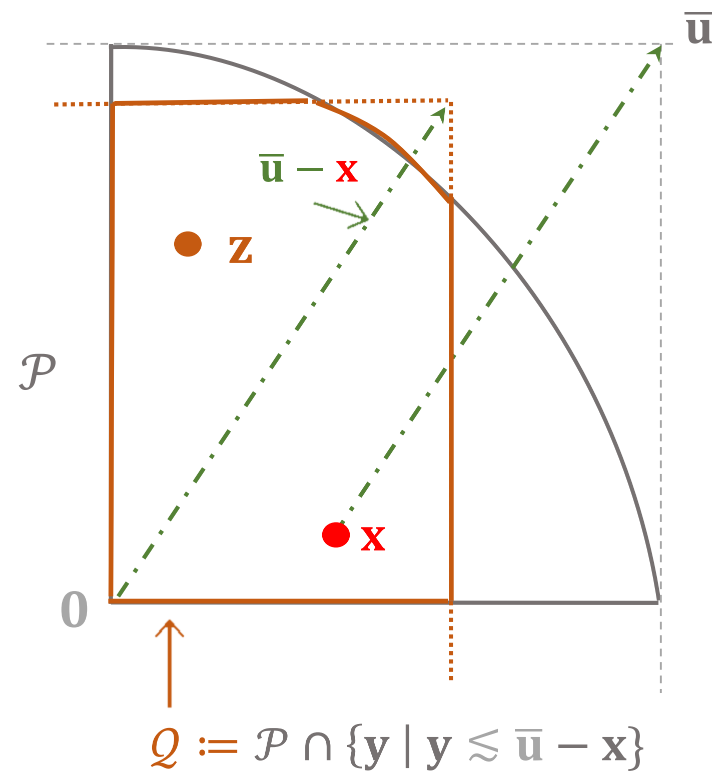
Local-Global relation
[BLKB '17]

$$\max_{\mathbf{x}\in\mathcal{P}} f(\mathbf{x})$$

- Let $\mathbf{x}$ be a stationary point in $\mathcal{P}$
- $\mathcal{Q} \coloneqq \mathcal{P} \cap \{\mathbf{y} \mid \mathbf{y} \lesssim \bar{\mathbf{u}} - \mathbf{x}\}$
- Let $\mathbf{z}$ be the a stationary point in $\mathcal{Q}$

Theorem:
$\max\{f(\mathbf{x}), f(\mathbf{z})\} \geq \frac{1}{4} f(\mathbf{x}^*)$

Can be generalized to approximately stationary points



Local-Global relation → Two-Phase algorithm

Two-Phase algorithm

$\mathbf{x} \leftarrow$ Non-convex Solver($\mathcal{P}$) // Phase I on $\mathcal{P}$

$\mathcal{Q} \leftarrow \mathcal{P} \cap \{\mathbf{y} \mid \mathbf{y} \lesssim \bar{\mathbf{u}} - \mathbf{x}\}$

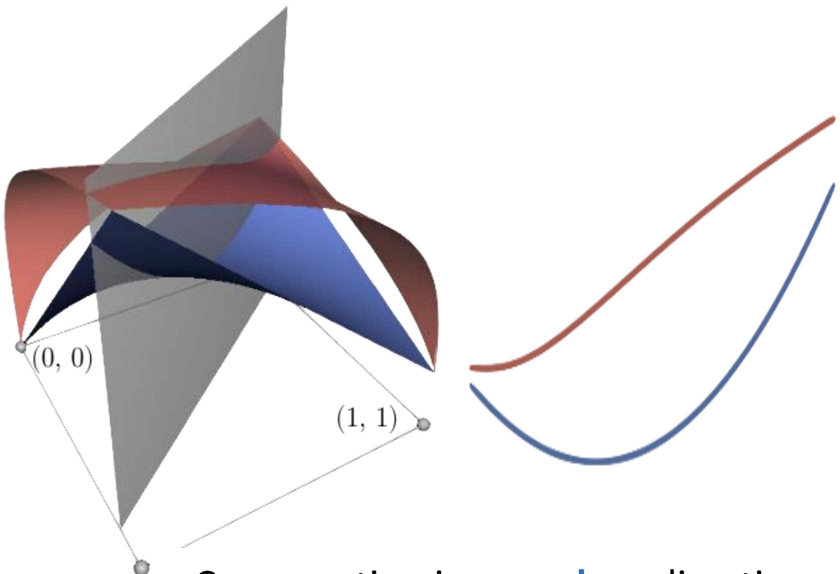
$\mathbf{z} \leftarrow$ Non-convex Solver($\mathcal{Q}$) // Phase II on $\mathcal{Q}$

Output: $\mathrm{argmax}\{f(\mathbf{x}), f(\mathbf{z})\}$

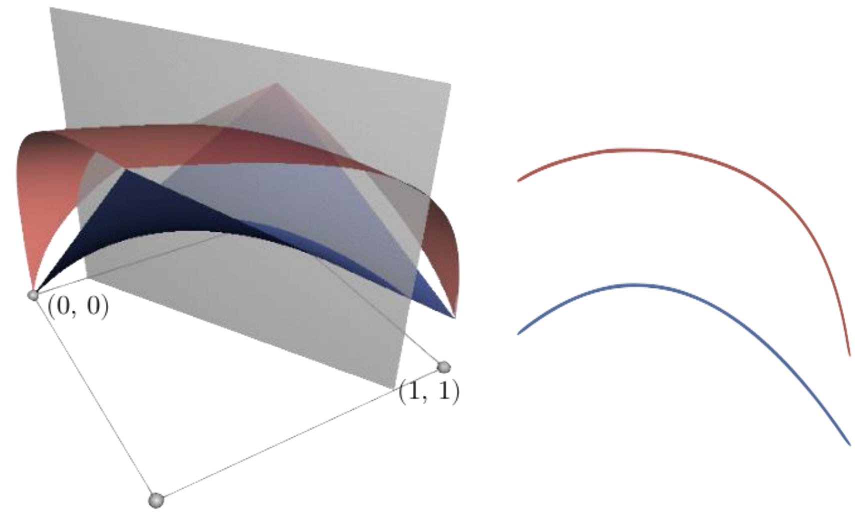
1/4 Guarantee

- Can use existing non-convex solvers to find (approximately) stationary points (used non-convex Frank-Wolfe in the experiments)
- Performs surprisingly good in experiments

Key property for a second algorithm



Cross-section in a **random** direction



Cross-section in a **positive** direction

Lemma: A DR-submodular f is concave along any non-negative direction.

Shrunken Frank-Wolfe: Follow concavity

[BLKB '17]

$$\max_{\mathbf{x} \in \mathcal{P}} f(\mathbf{x})$$

Shrunken FW

Choose initializer $\mathbf{x} \in \mathcal{P}$

In each iteration **do**:

Shrunken operator

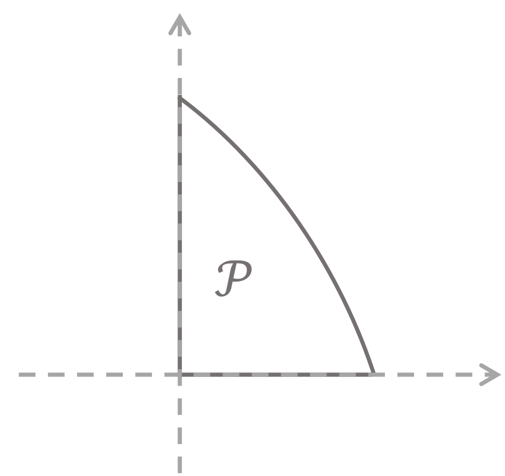
$$\mathbf{d} \leftarrow \text{argmax}_{\mathbf{v} \in \mathcal{P},\, \mathbf{v} \lesssim \bar{\mathbf{u}} - \mathbf{x}} \langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle$$

$$\mathbf{x} \leftarrow \mathbf{x} + \gamma \mathbf{d}$$

Return $\mathbf{x}$

Can make $\mathbf{d}$ to be a positive direction because:

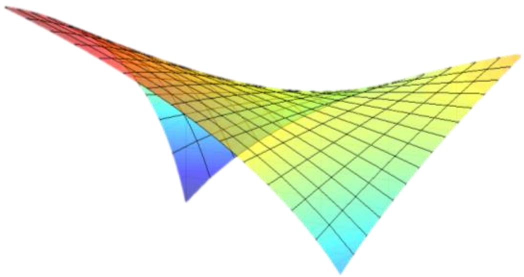
- $\mathbf{d}$ is from $\mathcal{P}$
- can always move $\mathcal{P}$ to the positive orthant without changing structure of the objective (since $\mathcal{P}$ is down-closed)



Theorem: $f(\mathbf{x}^K) \geq \frac{1}{e} f(\mathbf{x}^*) - \frac{LD^2}{2K}$

L: Lipschitz gradient, D: diameter of $\mathcal{P}$

Algorithm validation through the posterior agreement (PA) framework



Resulted in continuous DR-submodular maximization problems [BBK '19]

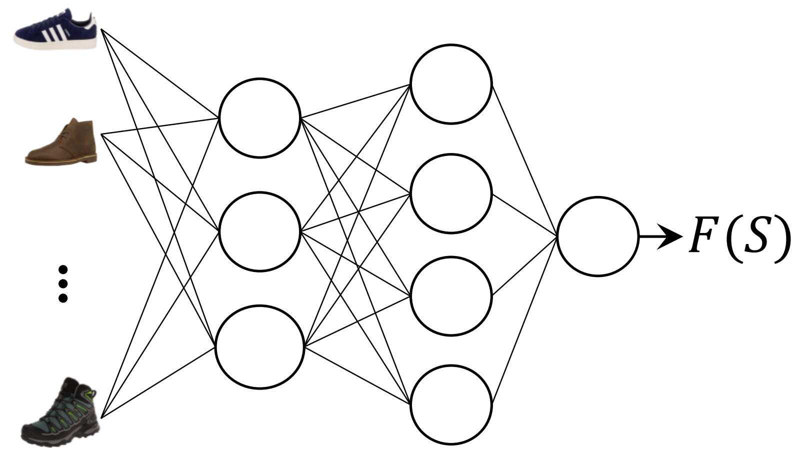
Motivation of posterior agreement

Product recommendation

Ground set $\mathcal{V}$: n products, n usually large

Which subset $S \subseteq \mathcal{V}$ to recommend?

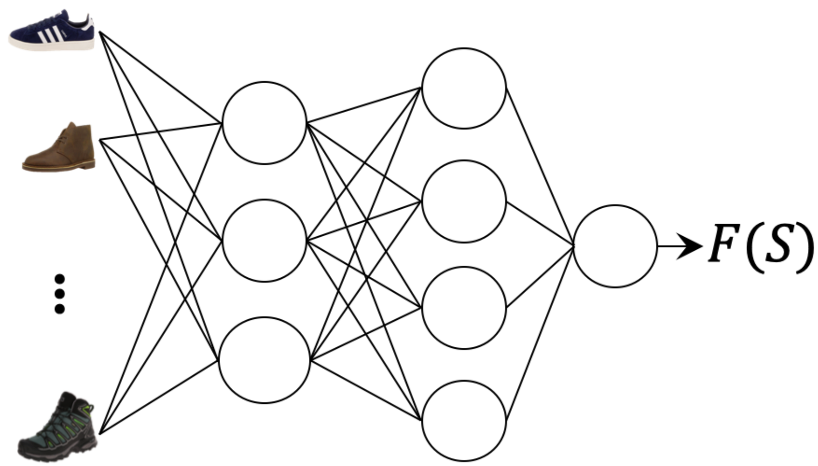
$F(S)$: a parameterized submodular utility function
e.g., a deep submodular neural net [Bilmes et al. '17]



Noisy training data D: a collection of chosen subsets by the users

- Learning: learn parameters and hyperparameters (architecture, stopping time & learning rate of SGD etc) of $F(S)$

- Inference: sample a subset from the distribution induced by $F(S)$



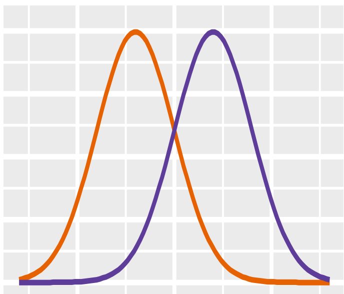
How to conduct inference and hyperparameter selection with noisy observations?

→ Can be achieved through the posterior agreement (PA) framework

Two-instance scenario, PA distribution and PA objective
[Buhmann '10, BGB '15, BGB '16]

hyperparameter choice

two noisy datasets		submodular utility functions		**posteriors** described by probabilistic log-submodular models [Djolonga et al. '14]
D'	$\rightarrow$	$F(S; D')$	$\rightarrow$	$p(S \mid D') \propto \exp(F(S; D'))$
D''	$\rightarrow$	$F(S; D'')$	$\rightarrow$	$p(S \mid D'') \propto \exp(F(S; D''))$



PA distribution:

$$p^{\mathrm{PA}}(S) \propto p(S \mid D')p(S \mid D'') \propto \exp[\, F(S; D') + (F(S; D'')\,]$$

used for inference

PA objective: measure the agreement between the two posteriors.
It is verified in an information-theoretic manner [BGB '16].

$$\sum_S p(S \mid D')p(S \mid D'')$$

used for hyperparameter validation

Inference via mean field approximation
[BBK '19]

Inference: sample from the PA distribution $p^{\mathrm{PA}}(S)$ → intractable

Mean field inference: approximate $p^{\mathrm{PA}}(S)$ by a factorized surrogate distribution:
$q(S|\mathbf{x}) := \prod_{i\in S} x_i \prod_{j\notin S}(1-x_j)$, $\mathbf{x} \in [0,1]^n$, then sample from $q(S|\mathbf{x})$

$$\log \mathrm{Z}^{\mathrm{PA}} = \log \textstyle\sum_S \exp[F(S;D') + (F(S;D'')] \qquad \text{(PA Evidence)}$$

$$\geq$$

$$\boxed{\mathbb{E}_{q(S|\mathbf{x})}[F(S;D')] + \mathbb{E}_{q(S|\mathbf{x})}[F(S;D'')] + \textstyle\sum_i H(x_i) =: f(\mathbf{x})} \qquad \text{(PA-ELBO)}$$

Provable mean field inference

[BBK '19]

$$\mathbb{E}_{q(S|\mathbf{x})}[F(S;D')] + \mathbb{E}_{q(S|\mathbf{x})}[F(S;D'')] + \sum_i H(x_i) =: f(\mathbf{x})$$ (PA-ELBO)

Finding a good lower bound → max (PA-ELBO) w.r.t. $q(S|\mathbf{x})$

Highly non-convex, however, Continuous DR-Submodular wrt $\mathbf{x}$ 😀

Box constrained continuous DR-submodular maximization problem:

$$\max f(\mathbf{x}), \text{ s.t. } \mathbf{x} \in [0,1]^n$$

Proposed a **tight ½** approximation algorithm: DR-DoubleGreedy

Validation of hyperparameters [BBK ‘19]

Given a hyperparameter choice,
PA objective measures how good the choice is.

PA objective is intractable → mean field inference provides lower bound

log $\sum_S p(S|D')p(S|D'')$ log PA objective

$\geq$

$\mathbb{E}_{q(S|\mathbf{x})}[F(S;D')] + \mathbb{E}_{q(S|\mathbf{x})}[F(S;D'')] + \sum_i H(x_i)$

(PA-ELBO) in last slide, provable algorithm from continuous DR-submodular maximization applies

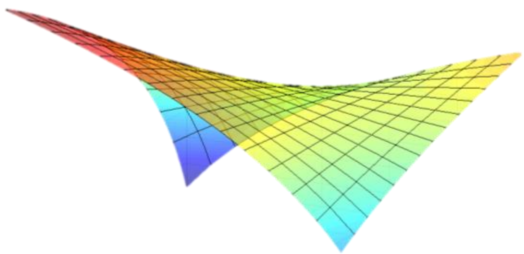
$-\ \log \sum_S \exp(F(S;\mathrm{D}'))$

$-\ \log \sum_S \exp F((S;D''))$

Two log partition functions, upper bounds exist [Djolonga et al. ‘14]

Experiments

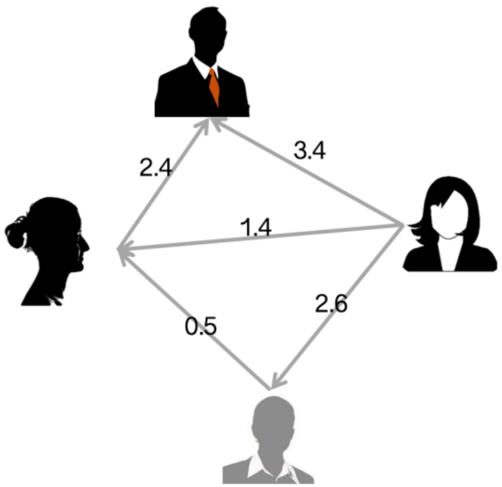
Revenue maximization



Revenue maximization: The details

W_{ij}: influence strength of i to j

Can be viewed as a variant of the Influence-and-Exploit strategy of [Hartline et al. '08].



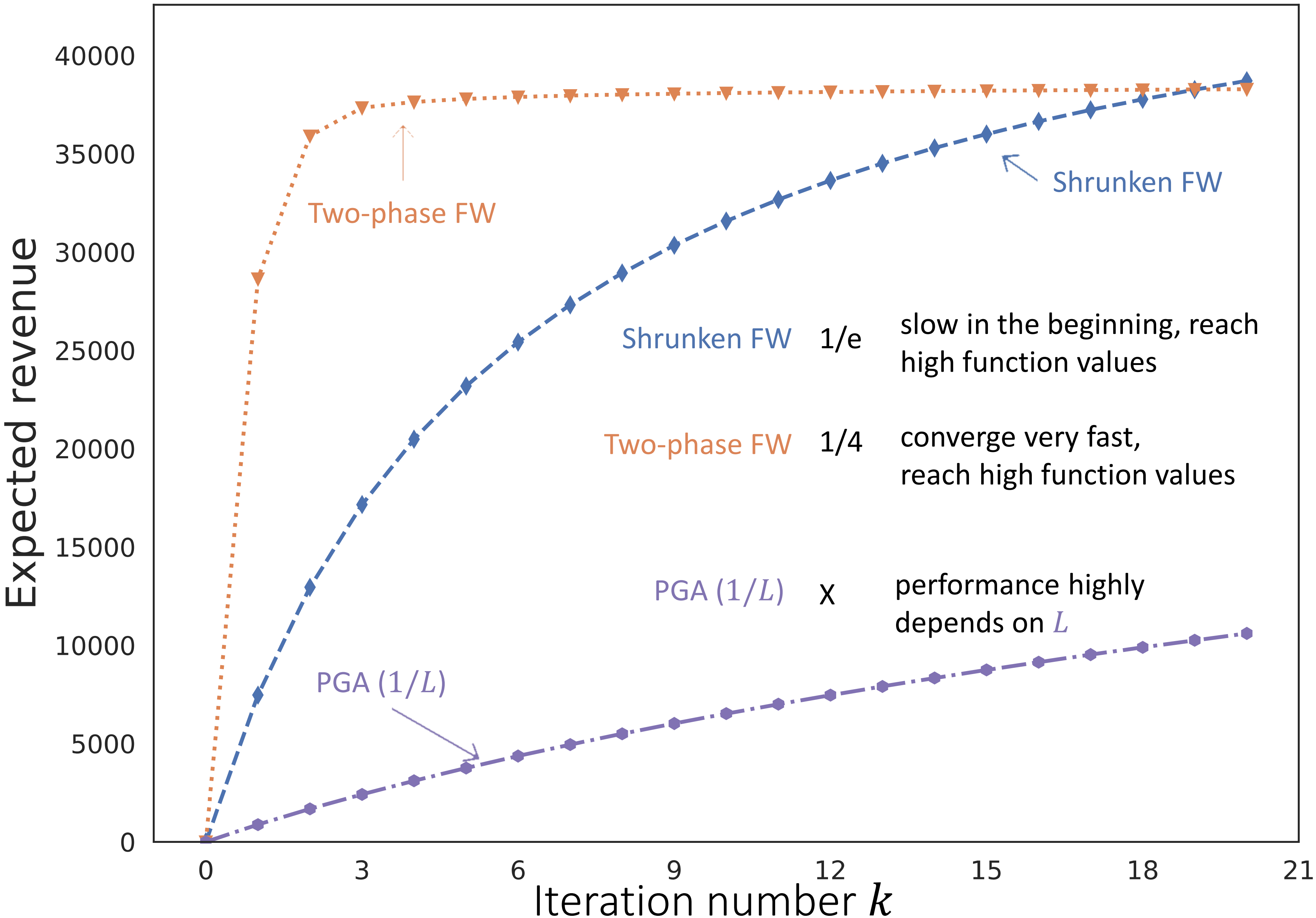
Influence stage: giving user i x_i units of products for free, he becomes an advocate with probability (independently from others) $1 - q^{x_i}$, $q \in (0,1)$ is a constant

Exploit stage: If a set S of users advocate the product, the resulted revenue is $R(S)$. The expected revenue is:

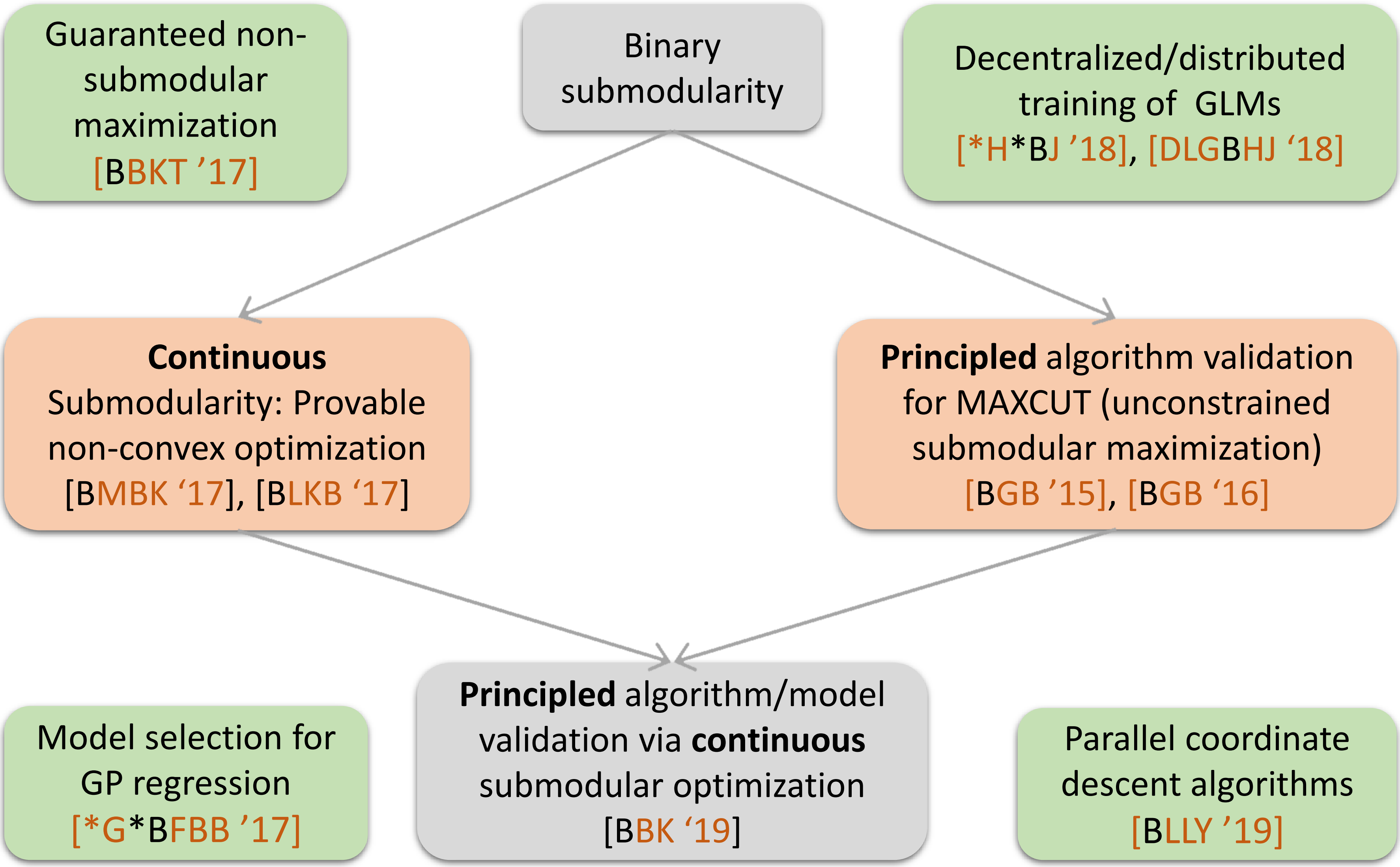
$$f(\mathbf{x}) = \mathbb{E}_S[R(S)] = \sum_{i \neq j} W_{ij}(1 - q^{x_i})q^{x_j}$$

Non-monotone DR-submodular

For simplicity, let $R(S)$ be the graph cut value of S



Results on "Ego Facebook" graph (4039 users), from the SNAP dataset


Guaranteed non-submodular maximization [BBKT ’17]
Binary submodularity
Decentralized/distributed training of GLMs [*H*BJ ’18], [DLGBHJ ‘18]
Continuous Submodularity: Provable non-convex optimization [BMBK ‘17], [BLKB ‘17]
Principled algorithm validation for MAXCUT (unconstrained submodular maximization) [BGB ’15], [BGB ‘16]
Model selection for GP regression [*G*BFBB ’17]
Principled algorithm/model validation via **continuous** submodular optimization [BBK ‘19]
Parallel coordinate descent algorithms [BLLY ’19]

Thanks for your attention!

Backup pages

Outlook

- Sampling methods for estimating the PA criterion

- Incorporate continuous submodularity as a prior knowledge into modern NN architecture
  - better generalization
  - more interpretable

- Explore submodularity over arbitrary conic lattices

Local-Global relation: Monotone setting

Strong relation between locally stationary points & global optimum [BLKB '17]

Lemma: for any two points $\mathbf{x}$, $\mathbf{y}$, it holds,
$(\mathbf{y} - \mathbf{x})^{\mathrm{T}} \nabla f(\mathbf{x}) \geq f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y}) - 2 f(\mathbf{x})$

Let $\mathbf{x}$ be a stationary point in $\mathcal{P}$ → e.g., $\nabla f(\mathbf{x}) = \mathbf{0}$

$$\max_{\mathbf{x} \in \mathcal{P}} f(\mathbf{x})$$

$\mathbf{a} \lesssim \mathbf{b}$ → $f(\mathbf{a}) \leq f(\mathbf{b})$

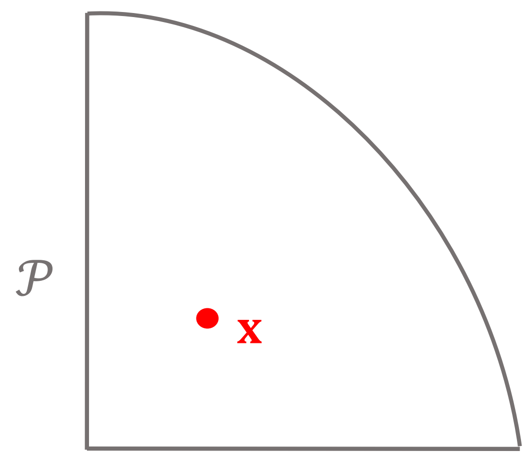
taking $\mathbf{y} = \mathbf{x}^*$ →

$$2 f(\mathbf{x}) \geq f(\mathbf{x} \vee \mathbf{x}^*) + f(\mathbf{x} \wedge \mathbf{x}^*)$$

$$\geq f(\mathbf{x}^*) + f(\mathbf{x} \wedge \mathbf{x}^*) \qquad \mathbf{x} \vee \mathbf{x}^* \gtrsim \mathbf{x}^*$$

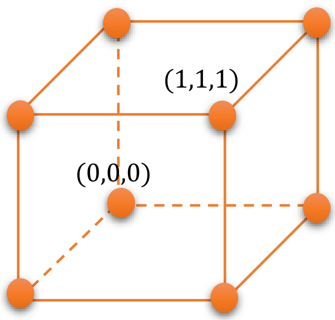
$$\geq f(\mathbf{x}^*) \qquad f(\mathbf{x} \wedge \mathbf{x}^*) \geq 0$$

$$f(\mathbf{x}) \geq \frac{1}{2} f(\mathbf{x}^*)$$



Generalized from submodularity of set functions

Ground set $\mathcal{V} = \{1, \ldots, n\}$ $F(X): 2^{\mathcal{V}} \mapsto \mathbb{R}_+$: utility, coverage, …

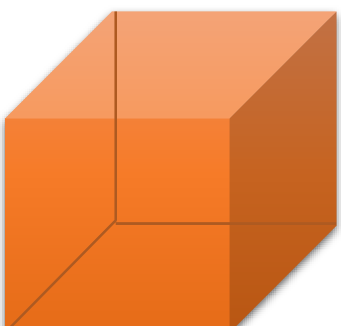


$$\forall\, X,\ Y \subseteq \mathcal{V},\quad F(X) + F(Y) \geq F(X \cup Y) + F(X \cap Y)$$

Equivalently, using binary vectors

$$\forall\, \mathbf{x},\ \mathbf{y} \in \{0,1\}^n,\ F(\mathbf{x}) + F(\mathbf{y}) \geq F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y})$$

$\vee$: coordinate-wise max. (JOIN)
$\wedge$: coordinate-wise min. (MEET)



$$\forall\, \mathbf{x},\ \mathbf{y} \in [a,\ b]^n,\ f(\mathbf{x}) + f(\mathbf{y}) \geq f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y})$$

Continuous submodularity (can be generalized to arbitrary lattice [Topkis ’78])

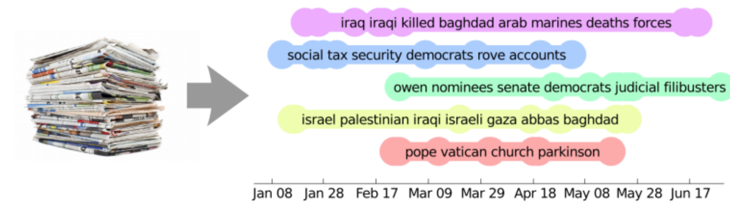
Supermodularity: f is supermodular iff $-f$ is submodular

MAP inference for DPPs [Gillenwater et al. '12]

DPP: determinantal point processes

- a distribution over subsets that favors diversity among items inside the subset
- originates from statistical physics [Macchi '75]

Task: Select a subset of points that are diverse



Softmax Extension for MAP inference:

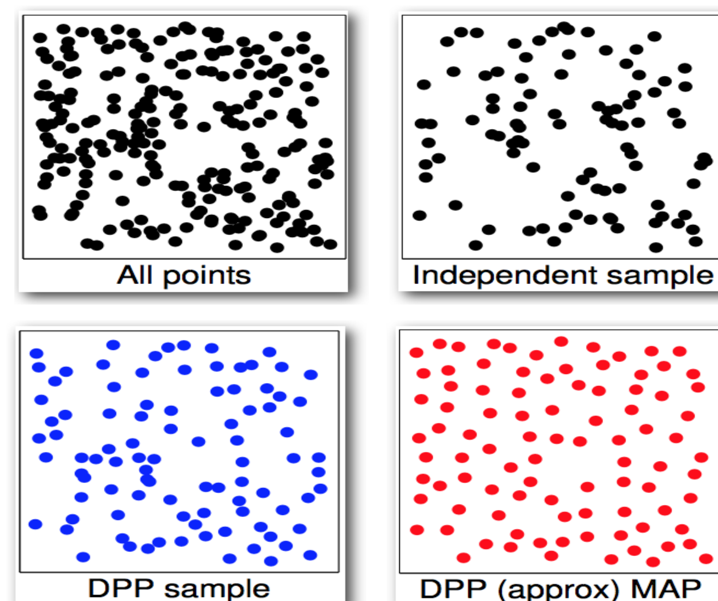
$$\mathrm{SE}(\mathbf{x}) = \log\det[\mathrm{diag}(\mathbf{x})(\mathbf{L} - \mathbf{I}) + \mathbf{I}]$$

$\mathbf{x} \in [0,1]^n$ ($x_i \rightarrow$ prob. of selecting item i)

$\mathbf{L}$: $n \times n$ kernel matrix, $\mathbf{L}_{ij}$: similarity between i, j

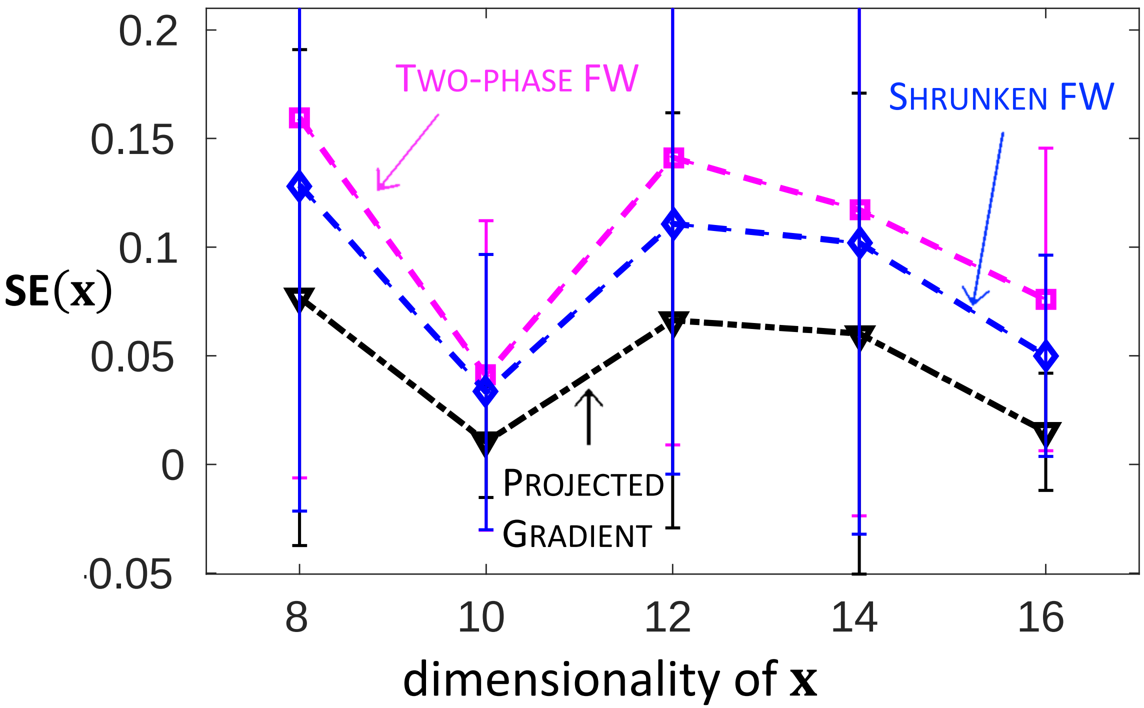
$\mathbf{I}$: identity matrix, $\mathrm{diag}(\mathbf{x})$: diagonal matrix

Proved continuous DR-submodularity → improved algorithm in both theory and practice



Experiments on the Softmax Extension: Synthetic

$\mathbf{SE}(\mathbf{x}) = \log\det[\mathrm{diag}(\mathbf{x})(\mathbf{L} - \mathbf{I}) + \mathbf{I}]$ Constraint: polytope



- Both better than PROJECTED GRADIENT

- TWO-PHASE FW $(\frac{1}{4})$ performs better than SHRUNKEN FW $(\frac{1}{e})$

Lessons:

- Sometimes worst-case analysis does not reflect practical performance
- More properties of $\mathbf{SE}(\mathbf{x})$ can be explored to explain practical performance

Real-world experiment: Matched summarization with DPPs [Gillenwater et al. '12]

Given statements made by A & B, select a set of pairs s.t. the two items *within* a pair are *similar*, but the set of pairs is *diverse.*

A1: No tax on interest, dividends, or capital gains. [tax]
A2: We're not going to have Sharia law applied in U.S. courts.
A3: I will ... grant a waiver from Obamacare to all 50 states. [Obamacare]
A4: We're spending more on foreign aid than we ought to. [foreign aid]
A5: If you think what we did in Massachusetts and what President Obama did are the same, boy, take a closer look. [Obamacare]

B1: I don't believe in a zero capital gains tax rate. [tax]
B2: Manufacture in America, you aren't going to pay any taxes. [tax]
B3: Zeroing out foreign aid ... that's absolutely the wrong course. [foreign aid]
B4: I voted against ethanol subsidies my entire time in Congress.
B5: Obamacare ... is going to blow a hole in the budget. [Obamacare]

Real-world experiment: Matched summarization with DPPs [Gillenwater et al. '12]

Given statements made by A & B, select a set of pairs s.t. the two items *within* a pair are *similar*, but the set of pairs are *diverse.*

A1: No tax on interest, dividends, or capital gains. [tax]

A3: I will ... grant a waiver from Obamacare to all 50 states. [Obamacare]
A4: We're spending more on foreign aid than we ought to. [foreign aid]

Can compare opinions of politicians on same topics

B1: I don't believe in a zero capital gains tax rate. [tax]

B3: Zeroing out foreign aid ... that's absolutely the wrong course. [foreign aid]

B5: Obamacare ... is going to blow a hole in the budget. [Obamacare]

Can be solved using DPP MAP inference with polytope constraints

Results on max. Softmax Extension $\mathbf{SE}(\mathbf{x})$

For TWO-PHASE FW, objectives of the selected phase were plotted.