Non-monotone Continuous DR-submodular Maximization: Structure and Algorithms

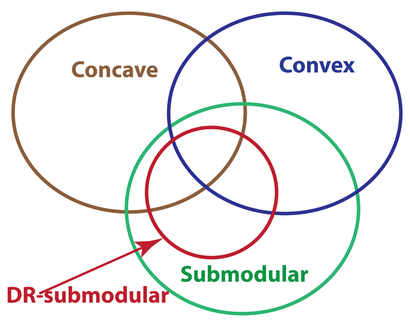
An Bian, Kfir Y. Levy, Andreas Krause and Joachim M. Buhmann

ETH zürich

Abstract

DR-submodularity captures a subclass of non-convex/non-concave functions that enables exact minimization and approximate maximization in poly. time.

- Investigate geometric properties that underlie such objectives, e.g., a strong relation between stationary points & global optimum is proved.
- Devise two guaranteed algorithms: i) A "two-phase" algorithm with ¼ approximation guarantee. ii) A non-monotone Frank-Wolfe variant with $^1/_e$ approximation guarantee
- Extend to a much broader class of submodular functions on "conic" lattices.



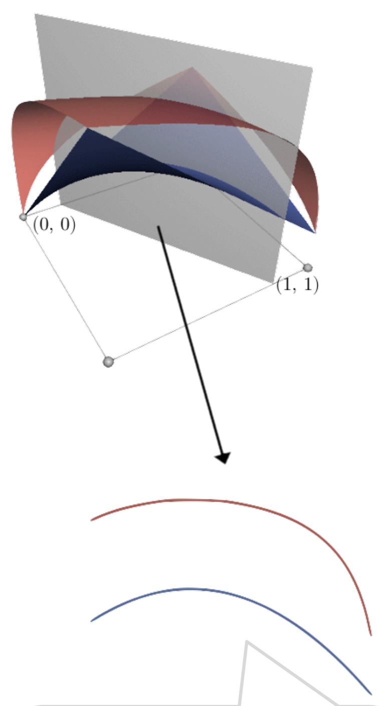
DR-submodular (Diminishing Returns) Maximization & Its Applications

$\max_{x \in \mathcal{P}} f(x)$ $f: \mathcal{X} \to \mathbb{R}$ is continuous DR-submodular. $\mathcal{X}$ is a hypercube. Wlog, let $\mathcal{X} = [\mathbf{0}, \bar{\mathbf{u}}]$. $\mathcal{P} \subseteq \mathcal{X}$ is convex and down-closed: $\mathbf{x} \in \mathcal{P}$ & $\mathbf{0} \le \mathbf{y} \le \mathbf{x}$ implies $\mathbf{y} \in \mathcal{P}$.

DR-submodular (DR property) [Bian et al '17]: $\forall\, \mathbf{a} \le \mathbf{b} \in \mathcal{X}, \forall\, i, \forall\, k \in \mathbb{R}_+$, it holds,

$$f(k\mathbf{e}_i + \mathbf{a}) - f(\mathbf{a}) \ge f(k\mathbf{e}_i + \mathbf{b}) - f(\mathbf{b}).$$

- If f differentiable, $\nabla f()$ is an antitone mapping ($\forall\, \mathbf{a} \le \mathbf{b}$, it holds $\nabla f(\mathbf{a}) \ge \nabla f(\mathbf{b})$)
- If f twice differentiable, $\nabla^2_{ij} f(\mathbf{x}) \le 0, \forall \mathbf{x}$



Softmax (red) & multilinear (blue) extensions, & concave cross-sections Fig. from [Gillenwater et al '12]

Applications

- Softmax extension for determinantal point processes (DPPs) [Gillenwater et al '12]
- Mean-field inference for log-submodular models [Djolonga et al '14]
- DR-submodular quadratic programming
- (Generalized submodularity over conic lattices) e.g., logistic regression with a non-convex separable regularizer [Antoniadis et al '11]
- Etc… (more see paper)

Underlying Properties of DR-submodular Maximization

Concavity Along Non-negative Directions:

Quadratic Lower Bound. With a L-Lipschitz gradient, for all $\mathbf{x}$ and $\mathbf{v} \in \pm\mathbb{R}^n_+$, it holds,

$$f(\mathbf{x}+\mathbf{v}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle - \tfrac{L}{2}\|\mathbf{v}\|^2$$

Strongly DR-submodular & Quadratic Upper Bound. f is μ-strongly DR-submodular if for all $\mathbf{x}$ and $\mathbf{v} \in \pm\mathbb{R}^n_+$, it holds,

$$f(\mathbf{x}+\mathbf{v}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle - \tfrac{\mu}{2}\|\mathbf{v}\|^2$$

Approximately Stationary Points & Global Optimum:

Lemma. For any $\mathbf{x}, \mathbf{y}$,

coordinate-wise max.   coordinate-wise min.

$$\langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{x}) \rangle \ge f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y}) - 2f(\mathbf{x}) + \tfrac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

If $\nabla f(\mathbf{x}) = 0$, then $2f(\mathbf{x}) \ge f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2 \to$ implicit relation between $\mathbf{x}$ & $\mathbf{y}$. (finding an exact stationary point is difficult 😯)

Non-stationarity Measure [Lacoste-Julien '16]. For any $\mathcal{Q} \subseteq \mathcal{X}$, the non-stationarity of $\mathbf{x} \in \mathcal{Q}$ is,

$$g_{\mathcal{Q}}(\mathbf{x}) := \max_{\mathbf{v} \in \mathcal{Q}} \langle \mathbf{v} - \mathbf{x}, \nabla f(\mathbf{x}) \rangle$$

(Local-Global Relation). Let $\mathbf{x} \in \mathcal{P}$ with non-stationarity $g_{\mathcal{P}}(\mathbf{x})$. Define $\mathcal{Q} := \mathcal{P} \cap \{\mathbf{y} \mid \mathbf{y} \le \bar{\mathbf{u}} - \mathbf{x}\}$. Let $\mathbf{z} \in \mathcal{Q}$ with non-stationarity $g_{\mathcal{Q}}(\mathbf{z})$. Then,

$$\max\{f(\mathbf{x}), f(\mathbf{z})\} \ge \tfrac{1}{4}[f(\mathbf{x}^*) - g_{\mathcal{P}}(\mathbf{x}) - g_{\mathcal{Q}}(\mathbf{z})] + \tfrac{\mu}{8}(\|\mathbf{x}-\mathbf{x}^*\|^2 + \|\mathbf{z}-\mathbf{z}^*\|^2),$$

where $\mathbf{z}^* := \mathbf{x} \vee \mathbf{x}^* - \mathbf{x}$.

- Proof using the essential **DR** property on carefully constructed auxiliary points
- Good empirical performance for the Two-Phase algorithm: if $\mathbf{x}$ is away from $\mathbf{x}^*$, $\|\mathbf{x}-\mathbf{x}^*\|^2$ will augment the bound; if $\mathbf{x}$ is close to $\mathbf{x}^*$, by the smoothness of f, should be near optimal.

Two Guaranteed Algorithms

NON-MONOTONE FRANK-WOLFE VARIANT

Input: step size $\gamma \in (0,1]$
$\mathbf{x}^{(0)} \leftarrow 0, k \leftarrow 0, t^{(0)} \leftarrow 0$ // t: cumulative step size
While $t^{(k)} < 1$ **do**:
   $\mathbf{v}^{(k)} \leftarrow \mathrm{argmax}_{\mathbf{v} \in \mathcal{P}, \mathbf{v} \le \bar{\mathbf{u}} - \mathbf{x}^{(k)}} \langle \mathbf{v}, \nabla f(\mathbf{x}^{(k)}) \rangle$ // shrunken LMO
   $\gamma_k \leftarrow \min\{\gamma, 1 - t^{(k)}\}$
   $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + \gamma_k \mathbf{v}^{(k)}$, $t^{(k+1)} \leftarrow t^{(k)} + \gamma_k$, $k{+}{+}$
Output: $\mathbf{x}^{(K)}$

key difference from the monotone Frank-Wolfe variant [Bian et al '17]

Guarantee of NON-MONOTONE FRANK-WOLFE VARIANT.

$$f(\mathbf{x}^{(K)}) \ge e^{-1} f(\mathbf{x}^*) - O\left(\tfrac{1}{K^2}\right) f(\mathbf{x}^*) - \tfrac{D^2 L}{2K}$$

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

Based on **Local-Global Relation**, can use any solver for finding an approximately stationary point as the subroutine, e.g., the Non-convex Frank-Wolfe solver in [Lascote-Julien '16]

TWO-PHASE ALGORITHM

Input: stopping tolerances ϵ_1, ϵ_2, #iterations K_1, K_2
$\mathbf{x} \leftarrow$ Non-convex Frank-Wolfe$(f, \mathcal{P}, K_1, \epsilon_1)$ // Phase I on $\mathcal{P}$
$\mathcal{Q} \leftarrow \mathcal{P} \cap \{\mathbf{y} \mid \mathbf{y} \le \bar{\mathbf{u}} - \mathbf{x}\}$
$\mathbf{z} \leftarrow$ Non-convex Frank-Wolfe$(f, \mathcal{Q}, K_2, \epsilon_2)$ // Phase II on $\mathcal{Q}$
Output: $\mathrm{argmax}\{f(\mathbf{x}), f(\mathbf{z})\}$

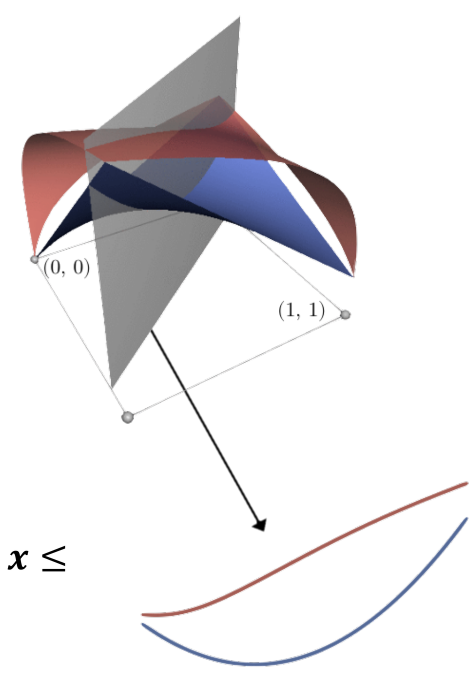
Guarantee of TWO-PHASE ALGORITHM.

$$\max\{f(\mathbf{x}), f(\mathbf{z})\} \ge \tfrac{\mu}{8}(\|\mathbf{x}-\mathbf{x}^*\|^2 + \|\mathbf{z}-\mathbf{z}^*\|^2) + \tfrac{1}{4}\left[f(\mathbf{x}^*) - \min\left\{\tfrac{C_1}{\sqrt{K_1+1}}, \epsilon_1\right\} - \min\left\{\tfrac{C_2}{\sqrt{K_2+1}}, \epsilon_2\right\}\right],$$

where $\mathbf{z}^* := \mathbf{x} \vee \mathbf{x}^* - \mathbf{x}$

Experimental Results (more see paper)

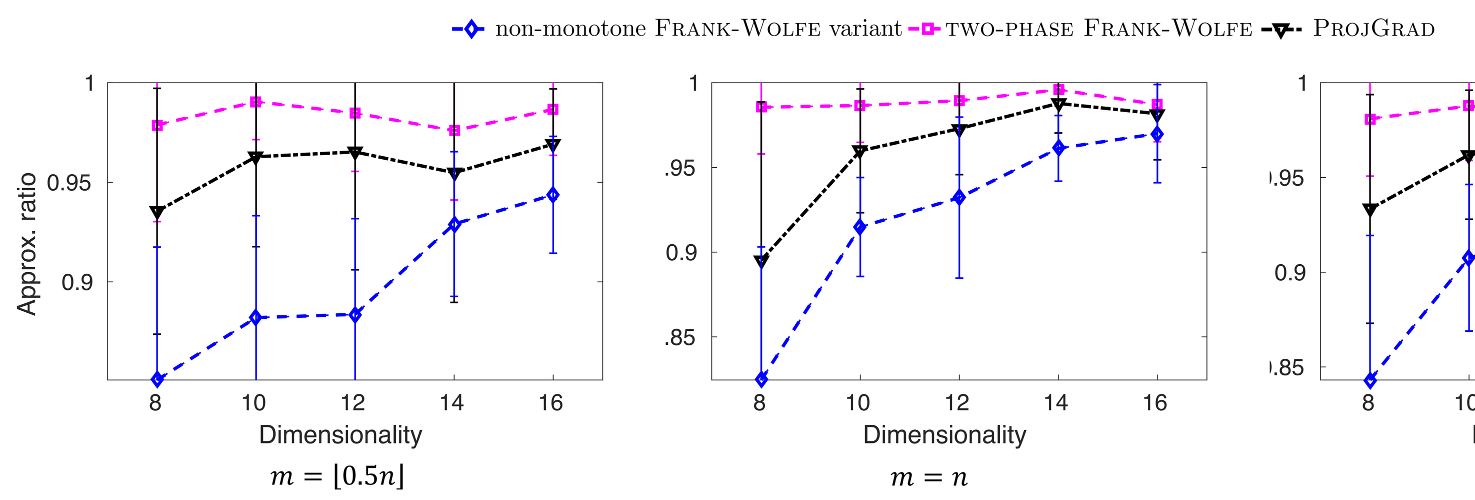
Baselines:
- QUADPROGIP: global solver for non-convex quadratic programming (possibly in exponential time)
- Projected Gradient Ascent (PROJGRAD) with diminishing step sizes ($^1/_{k+1}$)



DR-submodular Quadratic Programming.

Synthetic problem instances $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} + \mathbf{h}^\top \mathbf{x} + c$, $\mathcal{P} = \{\mathbf{x} \in \mathbb{R}^n_+ \mid \mathbf{A}\mathbf{x} \le \mathbf{b}, \mathbf{x} \le \bar{\mathbf{u}}, \mathbf{A} \in \mathbb{R}^{m \times n}_{++}, \mathbf{b} \in \mathbb{R}^m_+\}$ has m linear constraints.

Randomly generated in two manners:
1) Uniform distribution (see Figs below); 2) Exponential distribution

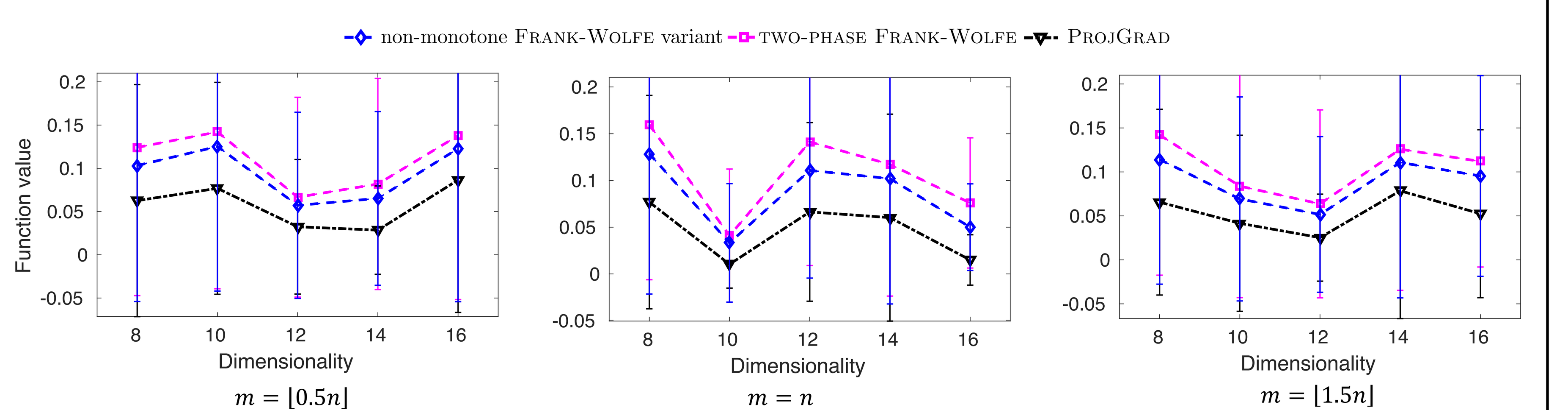


Maximizing Softmax Extensions for MAP inference of DPPs.

$$f(\mathbf{x}) = \log\det(\mathrm{diag}(\mathbf{x})(\mathbf{L} - \mathbf{I}) + \mathbf{I}), \mathbf{x} \in [0,1]^n$$

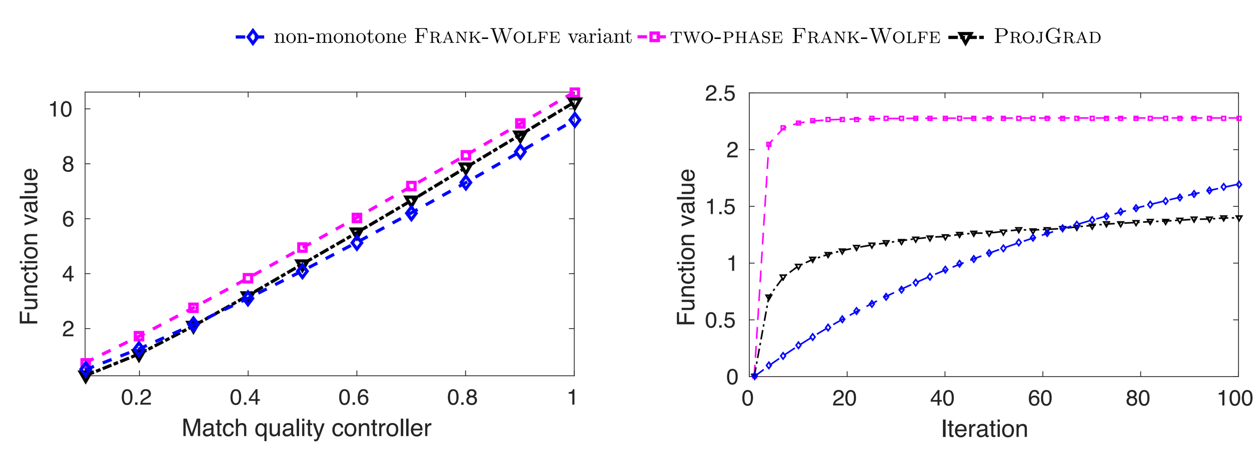
$\mathbf{L}$: kernel/similarity matrix. $\mathcal{P}$ is a matching polytope for matched summarization.

Synthetic problem instances:
- Softmax objectives: generate $\mathbf{L}$ with n random eigenvalues
- Generate polytope constraints similarly as that for quadratic programming



Real-world results on matched summarization:

Select a set of document pairs out of a corpus of documents, such that the two documents within a pair are similar, and the overall set of pairs is as diverse as possible. Setting similar to [Gillenwater et al '12], experimented on the 2012 US Republican detates data.



References

Feldman, Naor, and Schwartz. A unified continuous greedy algorithm for submodular maximization. FOCS 2011

Gillenwater, Kulesza, and Taskar. Near-optimal map inference for determinantal point processes. NIPS 2012.

Bach. Submodular functions: from discrete to continous domains. arXiv:1511.00394, 2015.

Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. arXiv:1607.00345, 2016.

Bian, Mirzasoleiman, Buhmann, and Krause. Guaranteed non-convex optimization: Submodular maximization over continuous domains. AISTATS 2017.