Optimal Continuous DR-Submodular Maximization and Applications to Provable Mean Field Inference

Yatao (An) Bian
Joachim M. Buhmann
Andreas Krause
ETH zürich

Example: Product recommendation: Men's shoes at Amazon

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

Which subset $S \subseteq \mathcal{V}$ to recommend? Need to:
1, learn a submodular utility function $\mathrm{F}(S)$ ➡ **Mean Filed Approximation Applies**
2, conduct approximate inference

Given a parameterized $\mathrm{F}(S)$ → Graphical model: $p(S) \propto e^{F(S)}$

Mean field aims to approximate $p(S)$ with a product distribution $q(S|\mathbf{x}) := \prod_{i\in S} x_i \prod_{j\notin S}(1-x_j), \mathbf{x}\in[0,1]^n$

$$\max_{\mathbf{x}\in[\mathbf{0},\mathbf{1}]} f(\mathbf{x}) := \overbrace{\mathbb{E}_{q(S|\mathbf{x})}[F(S)]}^{\text{multilinear extension of } F(S): f_{\text{mt}}(\mathbf{x})} - \sum_{i=1}^{n}[x_i \log x_i + (1-x_i)\log(1-x_i)]$$

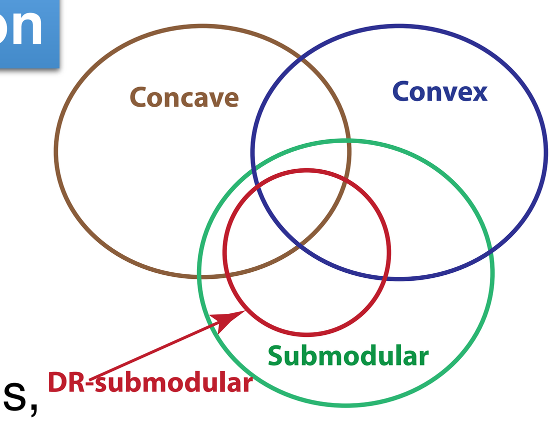
$$= f_{\text{mt}}(\mathbf{x}) + \sum_{i\in\mathcal{V}} H(x_i), \quad \text{(ELBO)}$$

😀 Continuous DR-Submodular wrt $\mathbf{x}$

Why mean field approximation?

1, Mean field as a differentiation technique → learn $F(S)$ end-to-end using modern deep learning framework
2, approximate inference using $q(S|\mathbf{x})$

Guaranteed Non-Convex Optimization: Continuous DR-Submodular (Diminishing Returns) Maximization



$$\underset{\mathbf{x}\in[\mathbf{a},\,\mathbf{b}]}{\text{maximize}} \quad f(\mathbf{x})$$ $f(\mathbf{x})$ is continuous DR-submodular

DR-submodular [Bian et al '17]: $\forall \mathbf{x} \leq \mathbf{y}, \forall i \in [n], \forall k \in \mathbb{R}_+$ it holds,

$$f(k\mathbf{e}_i + \mathbf{y}) - f(\mathbf{y}) \leq f(k\mathbf{e}_i + \mathbf{x}) - f(\mathbf{x})$$

= continuous submodularity (i.e. $f(\mathbf{x}) + f(\mathbf{y}) \geq f(\mathbf{x}\vee\mathbf{y}) + f(\mathbf{x}\wedge\mathbf{y})$)
+ coordinate-wise concavity

Continuous DR-Submodular Maximization is *NP-hard*. There is no $(\frac{1}{2}+\epsilon)$-approximation for any $\epsilon > 0$ unless RP=NP

Typical Applications

- Diversity models for recommendation [Tschiatschek et al '16, Djolonga et al '16]
- Data summarization [Lin et al '11]
- Model validation using posterior agreement [Bian et al '16]
- Variable selection [Krause et al '05]

Optimal Algorithm: DR-DoubleGreedy

Input: $\max_{\mathbf{x}\in[\mathbf{a},\mathbf{b}]} f(\mathbf{x}), \mathbf{x}\in\mathbb{R}^n$, $f(\mathbf{x})$ is DR-submodular

```
x^0 ← a, y^0 ← b;                                  (Maintain two solutions)
for k = 1 → n do
    let v_k be the coordinate being operated;
    find u_a such that f(x^{k-1}|_{v_k} u_a) ≥ max_{u'} f(x^{k-1}|_{v_k} u') − δ/n ;   } Solve 1-D problem on x
    δ_a ← f(x^{k-1}|_{v_k} u_a) − f(x^{k-1}) ;                                       }
    find u_b such that f(y^{k-1}|_{v_k} u_b) ≥ max_{u'} f(y^{k-1}|_{v_k} u') − δ/n ;   } Solve 1-D problem on y
    δ_b ← f(y^{k-1}|_{v_k} u_b) − f(y^{k-1}) ;                                       }
    x^k ← x^{k-1}|_{v_k}(δ_a/(δ_a+δ_b) u_a + δ_b/(δ_a+δ_b) u_b);   } Change coordinate to be a convex combination 😎
    y^k ← y^{k-1}|_{v_k}(δ_a/(δ_a+δ_b) u_a + δ_b/(δ_a+δ_b) u_b);   }
```

Output: $\mathbf{x}^n$ or $\mathbf{y}^n$ $(\mathbf{x}^n = \mathbf{y}^n)$

DR-DoubleGreedy has a 1/2-approximation guarantee
→ **Optimal** Algorithm

$$f(\mathbf{x}^n) \geq f(\mathbf{x}^*)/2 + [f(\mathbf{a}) + f(\mathbf{b})]/4 - 5\delta/4$$

δ: Error level in solving 1-D subproblem

Multi-epoch Extensions

```
Option I: DG-MeanField-1/3: run Submodular-DoubleGreedy to get a 1/3 initializer x̂
Option II: DG-MeanField-1/2: run DR-DoubleGreedy to get a 1/2 initializer x̂ ;
beginning with x̂, optimize f(x) coordinate by coordinate for T epochs ;
```

Experimental Results

One-epoch Algorithms	Multi-epoch Algorithms
- Submodular-DoubleGreedy (Sub-DG) [Bian et al '17a]	- CoordinateAscent-0 (0 as initializer)
- BSCB (Alg. 4 in [Niazadeh et al '18], optimal algorithm)	- CoordinateAscent-1
- DR-DoubleGreedy (DR-DG) (optimal one-epoch algorithm)	- CoordinateAscent-Random (random initializer)
	- BSCB-Multiepoch (Multi-epoch extension of BSCB)
	- DG-MeanField-1/3
	- DG-MeanField-1/2 (Multi-epoch extension of DR-DoubleGreedy)

FLID (facility location diversity model) [Tschiatschek et al '16]

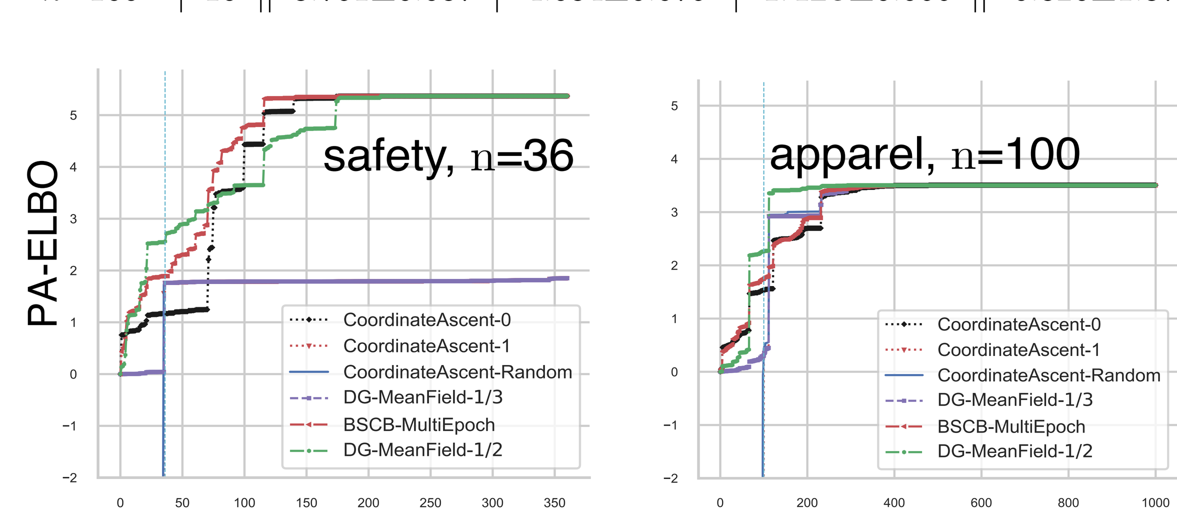
$$F(S) := \sum_{i\in S} u_i + \sum_{d=1}^{D}(\max_{i\in S} W_{i,d} - \sum_{i\in S} W_{i,d})$$

Statistics on one-epoch algorithms, boldface numbers indicate the best

Category	D	ELBO objective: Sub-DG	ELBO objective: BSCB	ELBO objective: DR-DG	PA-ELBO objective: Sub-DG	PA-ELBO objective: BSCB	PA-ELBO objective: DR-DG
carseats	2	2.089±0.166	2.863±0.090	**3.045**±0.069	1.015±1.081	2.106±0.228	**2.348**±0.219
	3	1.890±0.146	3.003±0.110	**3.138**±0.082	1.309±1.218	2.414±0.267	**2.707**±0.208
n=34	10	1.390±0.232	**3.100**±0.140	3.003±0.157	1.599±1.317	2.684±0.271	**2.915**±0.250
safety	2	1.934±0.402	2.727±0.212	**2.896**±0.098	1.370±1.203	2.049±0.280	**2.341**±0.161
	3	1.867±0.453	2.830±0.191	**2.970**±0.110	1.706±1.296	2.288±0.297	**2.619**±0.167
n=36	10	1.546±0.606	2.916±0.191	**2.920**±0.149	1.948±1.353	2.467±0.270	**2.738**±0.187
strollers	2	2.042±0.181	2.829±0.144	**2.928**±0.060	0.865±0.952	1.933±0.256	**2.202**±0.226
	3	1.814±0.264	2.958±0.146	**2.978**±0.077	1.172±1.063	2.181±0.297	**2.543**±0.254
n=40	10	1.328±0.544	**3.065**±0.162	2.910±0.140	1.702±1.334	2.480±0.304	**2.767**±0.336
media	2	3.221±0.066	3.309±0.055	**3.493**±0.051	0.372±0.286	**1.477**±0.128	1.336±0.101
	3	3.276±0.082	3.492±0.083	**3.712**±0.079	0.418±0.366	1.736±0.177	**1.762**±0.095
n=58	10	2.840±0.183	3.894±0.122	**3.924**±0.114	0.653±0.727	2.309±0.244	**2.524**±0.130
toys	2	3.543±0.047	3.454±0.091	**3.856**±0.044	0.597±0.480	1.731±0.182	**1.761**±0.133
	3	3.362±0.055	3.412±0.070	**3.736**±0.051	0.578±0.520	1.738±0.192	**1.802**±0.151
n=62	10	3.037±0.138	3.706±0.108	**3.859**±0.119	0.758±0.871	2.140±0.242	**2.330**±0.177
bedding	2	3.406±0.080	3.374±0.088	**3.620**±0.062	0.525±0.121	1.932±0.194	**2.001**±0.080
	3	3.648±0.106	3.564±0.083	**3.876**±0.081	2.499±0.972	2.250±0.269	**2.624**±0.066
n=100	10	3.355±0.161	3.799±0.144	**3.912**±0.082	**3.919**±0.045	2.578±0.358	3.157±0.091
apparel	2	3.560±0.094	3.527±0.046	**3.784**±0.059	0.268±0.109	**1.552**±0.141	1.513±0.191
	3	3.878±0.092	3.755±0.062	**4.140**±0.063	0.490±0.677	1.900±0.237	**2.225**±0.136
n=100	10	3.751±0.087	4.084±0.075	**4.425**±0.066	0.820±1.372	2.351±0.337	**2.967**±0.150

For ELBO, mean and standard deviation were calculated for 10 FLID models trained on 10 folds of the data, respectively

For PA-ELBO, mean and standard deviation were calculated for models trained over 45 pairs of folds



$$\max_{\mathbf{x}\in[0,1]^n} \mathbb{E}_{q(S|\mathbf{x})}[F(S|\mathrm{D}')] + \mathbb{E}_{q(S|\mathbf{x})}[F(S|\mathrm{D}'')] + \sum_{i\in\mathcal{V}} H(x_i) \quad \text{(PA-ELBO)}$$

PA (Posterior-Agreement) measures the agreement between two "noisy" posterior distributions

References

Bach. Submodular functions: from discrete to continuous domains. Mathematical Programming, 2018

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

Bian, Levy, Krause, Buhmann. Continuous DR-submodular Maximization: Structure and Algorithms. NIPS 2017b

Niazadeh, Roughgarden, Wang. Optimal Algorithms for Continuous Non-monotone Submodular and DR-Submodular Maximization. NIPS 2018

Tschiatschek, Djolonga, Krause. Learning Probabilistic Submodular Diversity Models Via Noise Contrastive Estimation. AISTATS 2016