Difference of submodular minimization via DC programming

Marwa El Halabi[1] George Orfanides[2] Tim Hoheisel[2]

[1]Samsung - SAIT AI Lab [2]McGill University

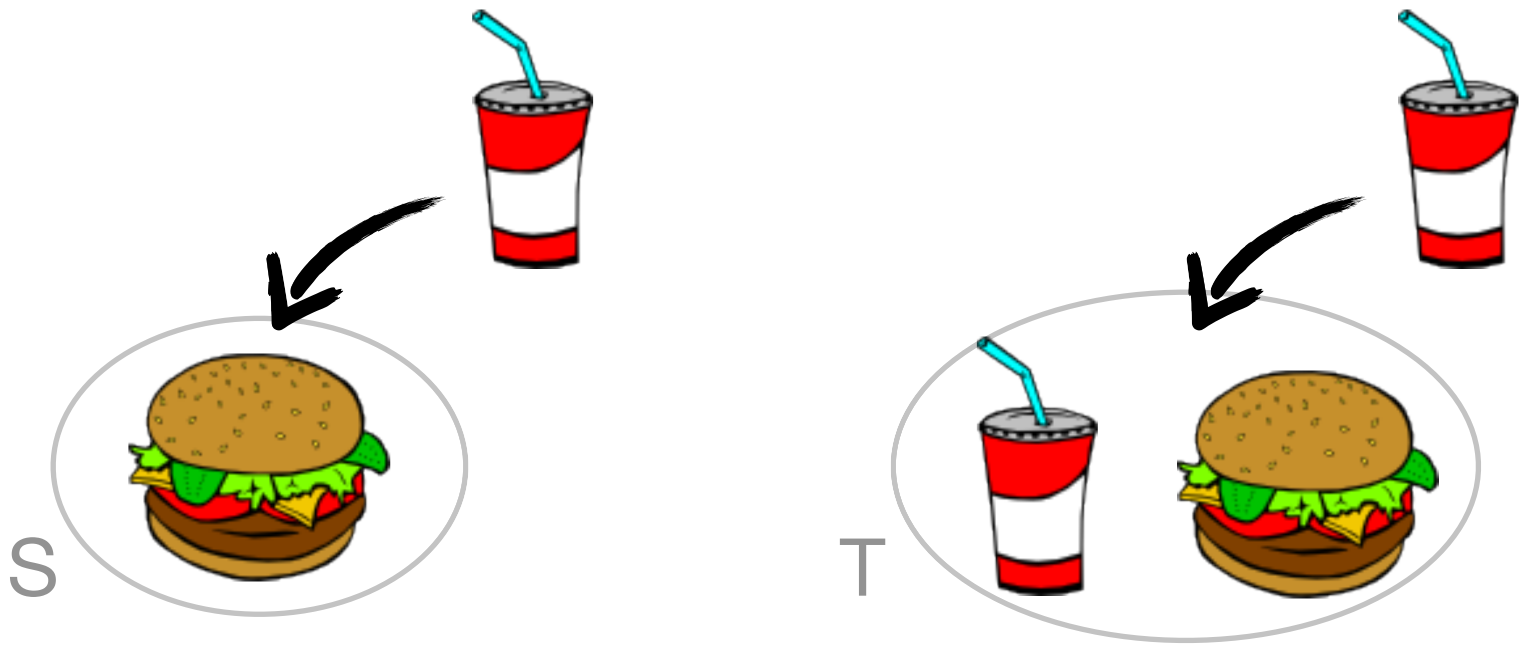
ICML 2023

Difference of submodular minimization

$$min_{S \subseteq V} F(S) := G(S) - H(S)$$

G and H are both:

- submodular:



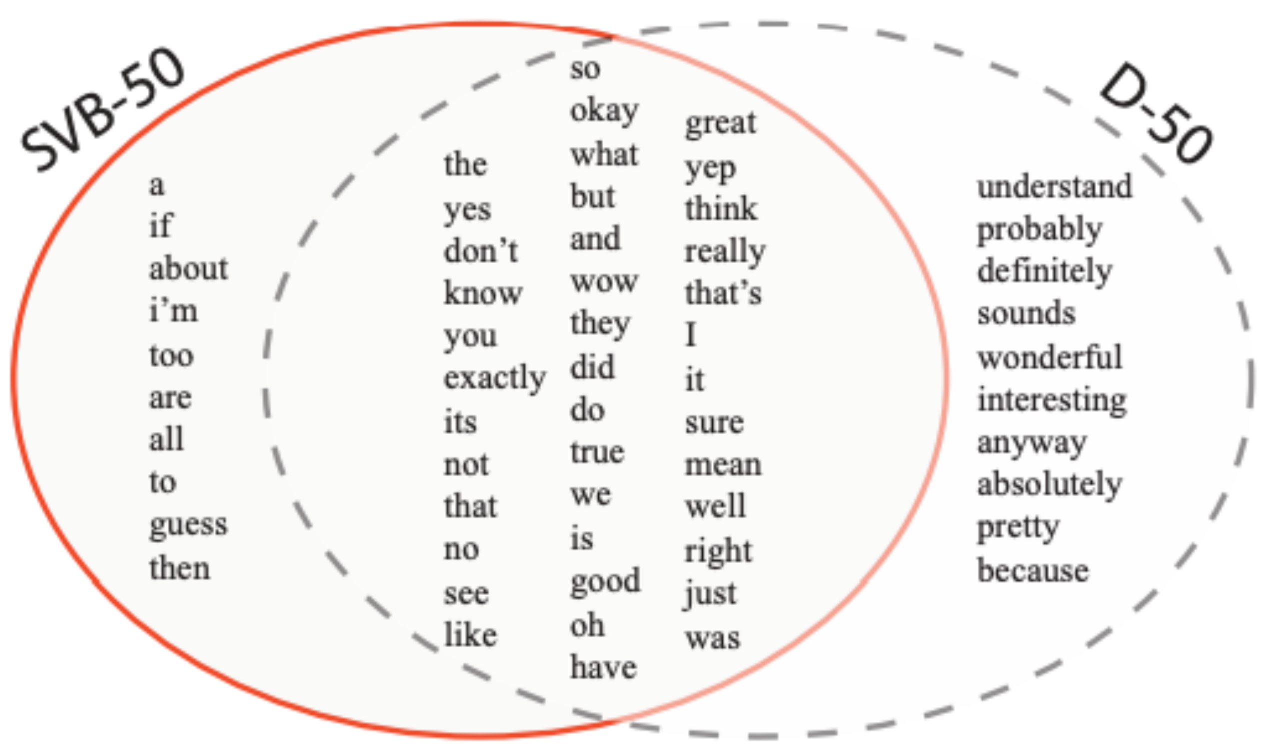
$$G(S \cup e) - G(S) \geq G(T \cup \{e\}) - G(T) \text{ for all } S \subseteq T$$

- normalized: $G(\emptyset) = 0$

Motivation

- Any set function can be expressed in this form (but computing it can be expensive) [Iyer & Bilmes, 2012].

- Many problems naturally have this form.



Speech corpus selection

Feature selection selection

- Very hard even if decomposition provided: No sub-exponential algorithm with constant factor approximation [Iyer & Bilmes, 2012]

Image sources: [Lin & Bilmes, 2011] and [Remeseiro & Bolon-Canedo, 2019]

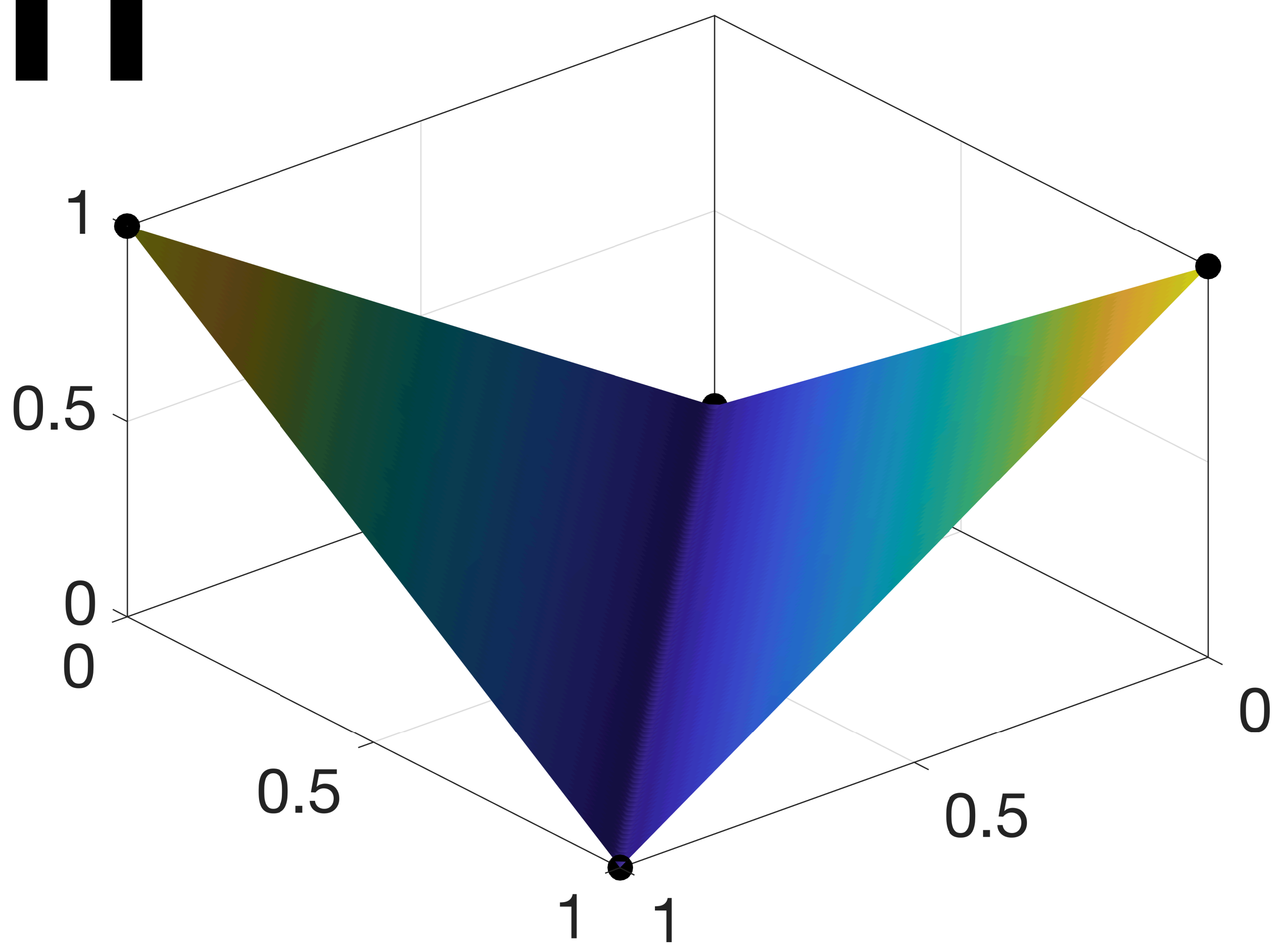
Existing approaches

- SubSup, SupSub, ModMod [Iyer & Bilmes, 2012]: converge to a local minimum

$$F(X) \leq F(X \cup i) \text{ for all } i \in V \backslash X, \text{ and } F(X) \leq F(X \backslash i) \text{ for all } i \in X.$$

- Projected gradient method (PGM) [El Halabi & Jegelka, 2020]: optimal approximation guarantee when F is approximately submodular.

Proposed approach



$$min_{S \subseteq V} F(S) := G(S) - H(S)$$

Solve equivalent continuous problem:

$$min_{S \subseteq V} F(S) = min_{x \in [0,1]^d} f_L(x) := g_L(x) - h_L(x) \quad (|V| = d)$$

- f_L, g_L, and h_L are the Lovász extensions of F, G, and H
- g_L and h_L are convex functions with easy to compute subgradients
- Given a minimizer x^* of f_L, we can obtain a minimizer of F by rounding

Proposed approach: DC Algorithm

$$min_{x\in[0,1]^d} f_L(x) := g_L(x) + \frac{\rho}{2}\|x\|^2 - (h_L(x) + \frac{\rho}{2}\|x\|^2)$$

DC algorithm (Pham Dinh & Le Thi, 1997):

$y^k \in \partial h_L(x^k) + \rho x^k$ (easy to compute)

$x^{k+1} \in \partial \left(g_L + \delta_{[0,1]^d} + \frac{\rho}{2}\|\cdot\|^2 \right)^* (y^k) = \underset{x\in[0,1]^d}{\text{argmin}}\, g_L(x) + \frac{\rho}{2}\|x - y^k/\rho\|^2$ (convex minimization)

Repeat until convergence $f_L(x^k) - f_L(x^{k+1}) \leq \epsilon$

If x^k is integral and $\rho = 0$: DCA is equivalent to SubSup

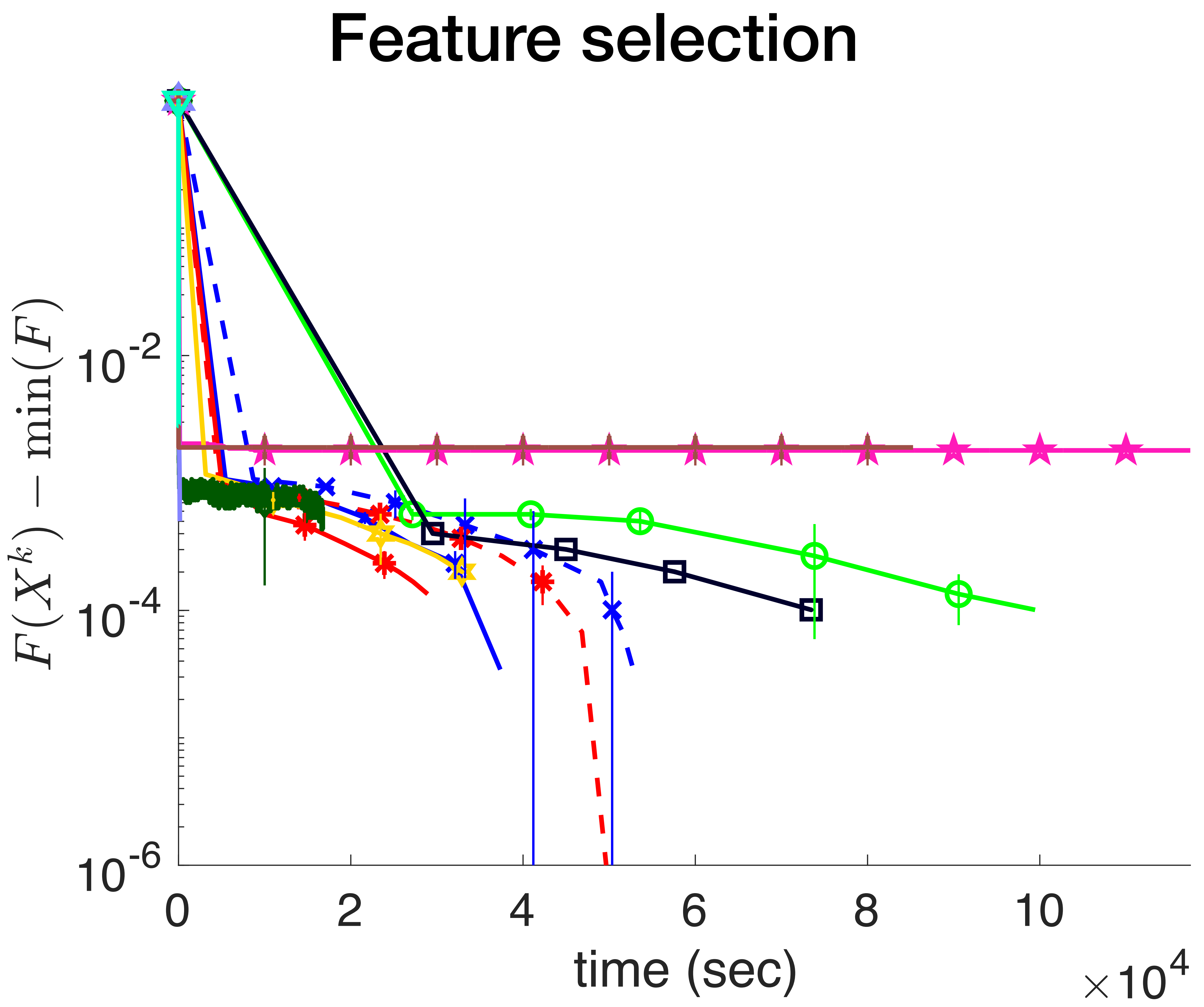
Theoretical guarantees

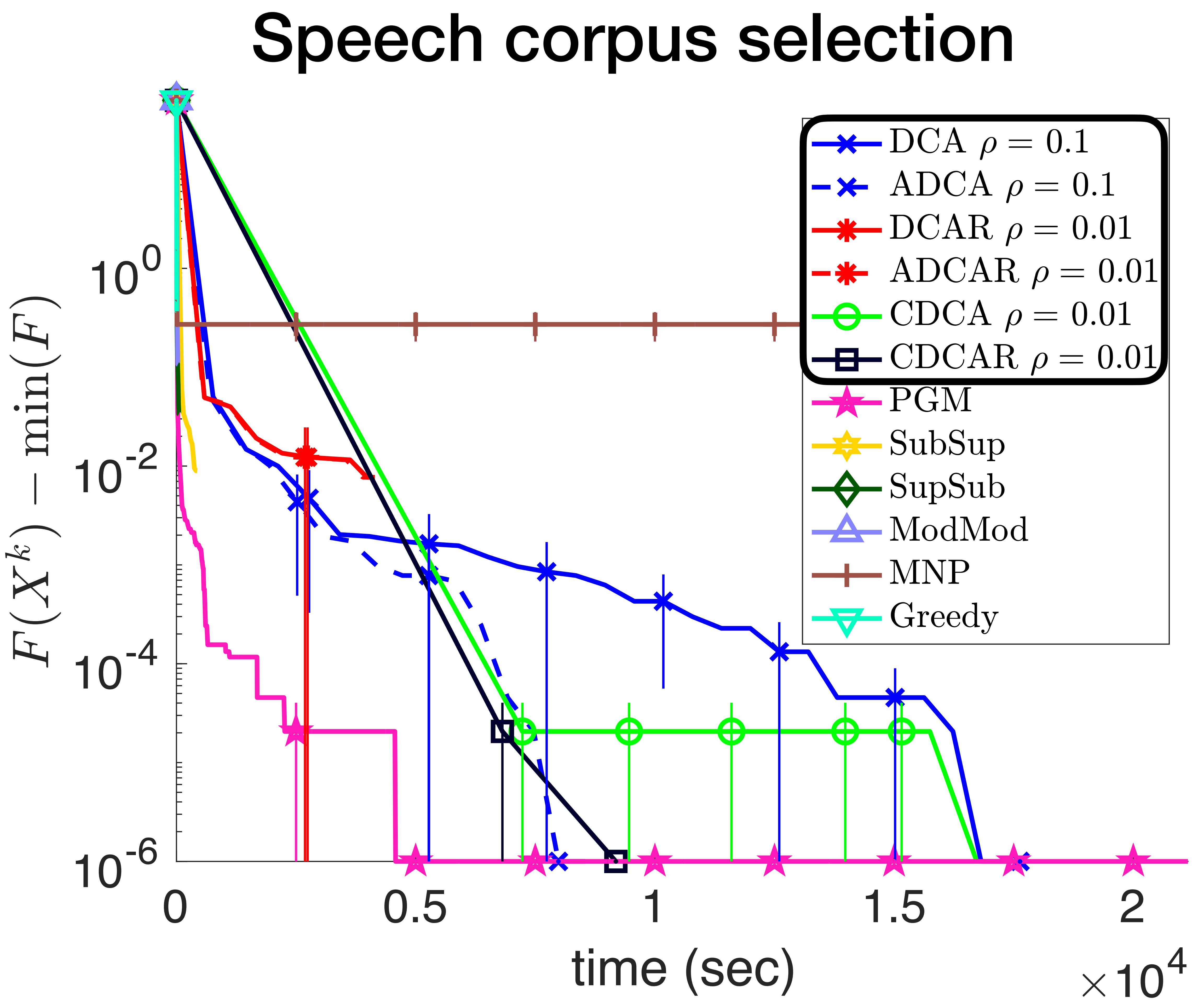
- If x^k is integral or we round at each iteration (necessary if $\rho > 0$):
  - DCA converges to a local minimum with rate $O(1/k)$ (similar to existing methods)
  - CDCA (variant of DCA) converges to a strong local minimum with rate $O(1/k)$ but requires solving a concave minimization (can obtain stationary point via FW)

for all $X' \subseteq X$ and $X' \supseteq X, F(X) \leq F(X')$

- **Effect of regularization:** convergence in iterates, slower convergence rate, faster convergence rate for subproblem

Empirical results





Special case: F is close to submodular; PGM, DCA and CDCA achieve optimal approximation guarantee here.

Thank you!

Paper: https://arxiv.org/abs/2305.11046

Code: https://github.com/samsungsailmontreal/difference-submodular-min

References

- El Halabi, M. and Jegelka, S. (2020). Optimal approximation for unconstrained non-submodular minimization. In International Conference on Machine Learning, pages 3961–3972. PMLR.

- Iyer, R. and Bilmes, J. (2012). Algorithms for approximate minimization of the difference between submodular functions, with applications. In Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI'12.

- Lin, H., & Bilmes, J. (2011). Optimal selection of limited vocabulary speech corpora. In Twelfth Annual Conference of the International Speech Communication Association.

- Pham Dinh, T. and Le Thi, H. A. (1997). Convex analysis approach to dc programming: theory, algorithms and applications. Acta mathematica vietnamica, 22(1):289–355.

- Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375.