Learning symmetric non-monotone submodular functions

Maria-Florina Balcan
Georgia Institute of Technology
ninamf@cc.gatech.edu

Nicholas J. A. Harvey
University of British Columbia
nickhar@cs.ubc.ca

Satoru Iwata
RIMS, Kyoto University
iwata@kurims.kyoto-u.ac.jp

Abstract

We prove a new structural result for symmetric submodular functions. We use that result to obtain an efficient algorithm for approximately learning such functions in the passive, supervised learning setting. We also complement this result with a nearly matching lower bound. Our work provides the first results for learning a large class of non-monotone submodular functions under general distributions.

1 Introduction

Submodular functions have historically played an important role in many areas, including combinatorial optimization, computer science, and economics [10]. More recently they have played an important role in algorithmic game theory and machine learning, e.g. [7, 8]. One reason for their ubiquity is that they have nice structural properties, in many ways similar to convexity of continuous functions, and these properties can be exploited algorithmically.

More recently there has been work on *learning submodular functions*. In particular, the paper of Goemans et al. [4] considers the problem of "approximate learning everywhere with value queries". For the class of *monotone* submodular functions, they give an algorithm which achieves an approximation factor $\tilde{O}(\sqrt{n})$, and they also show $\tilde{\Omega}(\sqrt{n})$ inapproximability. Their algorithm adaptively queries the target function at points of its choice, and produces a hypothesis that approximates the target function within a $\tilde{O}(\sqrt{n})$ at *every* point. A crucial step in their work is showing that any monotone, non-negative, submodular function can be approximated within a factor of $\sqrt{n}$ on every point by the square root of a linear function. Subsequent work of Balcan and Harvey [1] use this result to provide an algorithm for learning monotone submodular functions with a $\tilde{O}(\sqrt{n})$ approximation factor, in the distributional (passive supervised) learning setting. Here the goal is to design an efficient algorithm which outputs a hypothesis that provides a multiplicative approximation of the target function on most of the samples coming from an underlying data distribution, given only a polynomial number of i.i.d. samples from that distribution.

In this paper we consider *symmetric* submodular functions. These are a natural class of submodular functions to consider, as they generalize cut capacity functions of graphs, and they have played an important role in both the structural theory and algorithmic theory of submodular functions [3, 9]. We prove a new structural result for symmetric submodular functions: they can be approximated within a factor of $\sqrt{n}$ on every point by the square root of a quadratic function. We use this result to provide a polynomial time algorithm for learning such functions with an approximation factor of $O(\sqrt{n})$ in the approximate distributional learning setting of Balcan and Harvey [1]. In particular, we show how to reduce the problem of learning a submodular function in a distributional learning setting to the problem learning a linear separator in the usual PAC model in an appropriately constructed

feature space. By providing an approximation preserving reduction from the problem of learning monotone submodular functions to the problem of learning symmetric submodular functions, we additionally show that symmetric submodular functions cannot be learned in this model with an approximation factor $\tilde{o}(n^{1/3})$.

Observe that a symmetric submodular function is not monotone unless it is constant. Thus, our work provides the first results for learning a large class of non-monotone submodular functions under general distributions.

Related Work A series of recent papers have considered the problem of learning submodular functions in a distributional learning setting. In addition to results on approximate learning everywhere with value queries [4] or on approximate distributional learning on that apply on arbitrary input distributions, there have also been results specific to product distributions.

Balcan and Harvey [1] provide an algorithm for PMAC-learning monotone Lipschitz submodular functions under product distributions with a constant approximation factor. In subsequent work, building on a technique of [1], Gupta et al. [5] present an algorithm for learning non-monotone submodular functions under a product distribution based on value queries. Their guarantee is additive (assuming the function is appropriately normalized), rather than multiplicative and their running time is $n^{\mathrm{poly}\,(1/\epsilon)}$. Cheraghchi et al. [2] study the noise stability of submodular functions. As a consequence they obtain an algorithm for learning a submodular function under product distributions. Their algorithm also works for non-monotone and non-Lipschitz functions, and only requires access to the submodular function via statistical queries and its running time is $n^{\mathrm{poly}\,(1/\epsilon)}$. Their algorithm is agnostic (meaning that they do not assume the target function is submodular), and their performance guarantee proves that the L_1-loss of their hypothesis is at most ϵ more than the best error achieved by any submodular function (assuming the function is appropriately normalized). Note however, that as opposed to our work, the Cheraghchi et al. [2] and Gupta et al. [5] algorithms work only for product distributions and their running time is $n^{\mathrm{poly}\,(1/\epsilon)}$ as opposed to $\mathrm{poly}\,(n, 1/\epsilon)$.

2 Preliminaries

Let $[n] = \{1, \ldots, n\}$. A function $f : 2^{[n]} \to \mathbb{R}$ is

- *Normalized* if $f(\emptyset) = 0$.
- *Non-negative* if $f(S) \geq 0$ for all S.
- *Monotone* (or *non-decreasing*) if $f(S) \leq f(T)$ for all $S \subseteq T$.
- *Symmetric* if $f(S) = f([n] \setminus S)$ for all S.
- *Submodular* if it satisfies

$$f(S+i+j) - f(S+j) \;\leq\; f(S+i) - f(S) \qquad \forall S \subseteq [n],\, i, j \in [n], \tag{1}$$

$$\text{or equivalently} \qquad f(S) + f(T) \;\geq\; f(S \cup T) + f(S \cap T) \qquad \forall S, T \subseteq [n]. \tag{2}$$

For the rest of this paper we will implicitly assume that all submodular functions are normalized.

For a subset $U \subseteq [n]$, let $\chi(U) \in \{0,1\}^n$ denote the characteristic vector of U, i.e., $\chi(U)_i = 1$ iff $i \in U$. For a vector $x \in \mathbb{R}^n$ and a set $U \subseteq [n]$, let $x(U) = \chi(U)^{\mathsf{T}} x = \sum_{i \in U} x_i$.

Let f be a submodular function. The *extended polymatroid* [10, Eq. (44.5)] associated with f is

$$EP_f \;:=\; \{\, x \in \mathbb{R}^n \,:\, x(U) \leq f(U) \;\; \forall U \subseteq [n] \,\}.$$

The *base polytope* [10, Eq. (44.7)] of f is a facet of EP_f, defined by

$$B_f \;:=\; EP_f \,\cap\, \{\, x \in \mathbb{R}^n \,:\, x([n]) = f([n]) \,\}.$$

A set $Q \subseteq \mathbb{R}^n$ is called centrally symmetric if $x \in Q \iff -x \in Q$. Given a symmetric, positive definite matrix A of size $n \times n$, let $E(A)$ denote the ellipsoid centered at the origin defined by A:

$$E(A) \;=\; \{\, x \in \mathbb{R}^n : x^{\mathsf{T}} A x \leq 1 \,\}.$$

3 Structural results

The paper of Goemans et al. [4] showed that any monotone submodular function can be approximated within a factor of $\sqrt{n}$ on every point by the square root of a linear function. The proof is based on interesting tools from convex geometry, such as John's ellipsoid theorem, which shows that any centrally symmetric convex body $Q \subset \mathbb{R}^n$ can be approximated to within a factor of $\sqrt{n}$ by an ellipsoid E (known as the John ellipsoid). (Formally, $E \subseteq Q \subseteq \sqrt{n} \cdot E$.) The base polytope of a monotone submodular function f is always contained within the non-negative orthant, so it is not centrally symmetric, but one can easily symmetrize it by reflecting it into all orthants and taking the convex hull. Goemans et al. show that optimizing $\{0,1\}$-linear functions over the resulting body recovers the function f, and therefore optimizing over that body's John ellipsoid instead approximates f to within a factor $\sqrt{n}$. Furthermore, that ellipsoid is shown to be axis-aligned, so optimizing over it in the direction $c \in \{0,1\}^n$ amounts to evaluating the square-root of a linear function in c.

This approach fails for *non-monotone* submodular functions because their base polytope is not contained within the non-negative orthant, so reflecting it and taking the convex hull ruins the structure. Our main observation is that, for *symmetric* submodular functions, the base polytope is already centrally symmetric, so John's ellipsoid theorem can be applied directly to that polytope, without performing any reflections. In this case the John ellipsoid is not axis-aligned, so optimizing over that ellipsoid produces the square-root of a *quadratic* function, rather than a linear function. As we show in the following section this is enough to design a learning algorithm with provable guarantees in a distributional setting.

Lemma 1. Let $f : 2^{[n]} \to \mathbb{R}$ be a symmetric, submodular function. Then:

(1): B_f is centrally symmetric, and

(2): $f(S) = \max\left\{ \chi(S)^{\mathsf{T}} x \,:\, x \in B_f \right\}$ for all $S \subseteq [n]$.

Proof. Since f is normalized and symmetric, we have $f([n]) = f(\emptyset) = 0$. Thus

$$B_f \;=\; EP_f \;\cap\; \left\{ x \,:\, \sum_i x_i = 0 \right\}.$$

Fix any $x \in B_f$. Then, for any $U \subseteq [n]$,

$$-x(U) \;=\; -\sum_{i \in U} x_i \;=\; \underbrace{\sum_{i=1}^{n} x_i}_{=0} - \sum_{i \in U} x_i \;=\; \sum_{i \in [n] \setminus U} x_i \;\leq\; f([n] \setminus U) \;=\; f(U),$$

where the inequality comes from $x \in EP_f$ and the last equality holds since f is symmetric. This shows that $-x \in EP_f$, and since since $-\sum_{i=1}^n x_i = 0$, we also have $-x \in B_f$. This proves (1).

To prove (2), we will use the greedy algorithm. Fix any $S \subseteq [n]$. Let $\pi : [n] \to [n]$ be an ordering of $[n]$ such that $\pi(i) \in S$ for all $i = 1, \ldots, |S|$. Define $U_0, \ldots, U_n \subseteq [n]$ and $x \in \mathbb{R}^n$ by

$$\begin{aligned} U_i \;&=\; \{\pi(1), \ldots, \pi(i)\} \\ x_{\pi(i)} \;&=\; f(U_i) - f(U_{i-1}). \end{aligned}$$

It is known [10, Theorem 44.3] that x is an optimal solution of $\max\left\{ \chi(S)^{\mathsf{T}} x \,:\, x \in EP_f \right\}$. In fact, it is also an optimal solution when optimizing over B_f because $\sum_{i=1}^n x_i = f([n]) - f(\emptyset) = f([n])$, by telescoping and since f is normalized. Furthermore,

$$\chi(S)^{\mathsf{T}} x \;=\; \sum_{i \in S} x_i \;=\; \sum_{1 \leq i \leq |S|} x_{\pi(i)} \;=\; f(S) - f(\emptyset) \;=\; f(S)$$

proving (2). ■

Theorem 2. Let $f : 2^{[n]} \to \mathbb{R}_+$ be a symmetric submodular function. Then there exists a function $\hat{f}$ of the form $\hat{f}(S) = \sqrt{\chi(S)^{\mathsf{T}} M \chi(S)}$ where M is a symmetric, positive definite matrix, such that $\hat{f}(S) \leq f(S) \leq \sqrt{n}\hat{f}(S)$ for all $S \subseteq [n]$.

Proof. Following Goemans et al. [4], we use John's ellipsoid theorem to show that there exists an ellipsoid E centered at the origin such that

$$E \subseteq B_f \subseteq \sqrt{n}E.$$

This holds because of Lemma 1, claim (1). Then for any $S \subseteq [n]$ we have

$$\max\left\{ \chi(S)^{\mathsf{T}}x : x \in E \right\} \leq \max\left\{ \chi(S)^{\mathsf{T}}x : x \in K \right\} \leq \max\left\{ \chi(S)^{\mathsf{T}}x : x \in \sqrt{n}E \right\}. \tag{3}$$

Define $\hat{f}(S) = \max\left\{ \chi(S)^{\mathsf{T}}x : x \in E \right\}$. Then Lemma 1, claim (2) implies that

$$\hat{f}(S) \leq f(S) \leq \sqrt{n} \cdot \hat{f}(S) \qquad \forall S \subseteq [n],$$

so f is approximated everywhere by $\hat{f}$ to within a factor $\sqrt{n}$.

Given any non-zero $c \in \mathbb{R}^n$, we have that

$$\begin{aligned} \max\{\, c^{\mathsf{T}}x : x \in E(A) \,\} &= \max\{\, c^{\mathsf{T}}A^{-1/2}x : \|x\| \leq 1 \,\} \\ &= c^{\mathsf{T}}A^{-1/2}\Big(\frac{A^{-1/2}c}{\|A^{-1/2}c\|}\Big) = \sqrt{c^{\mathsf{T}}A^{-1}c} \end{aligned}$$

Thus $\hat{f}(S)$ has the closed-form expression $\hat{f}(S) = \sqrt{\chi(S)^{\mathsf{T}}A^{-1}\chi(S)}$, as desired. ■

We emphasize that, for monotone functions the relevant ellipsoid is axis-aligned, so the matrix A is diagonal and $\hat{f}(S)$ is the square root of a linear function in S. For symmetric functions, A will typically not be diagonal, and therefore the function $\hat{f}(S)$ is indeed quadratic in S.

4 PMAC learning of symmetric submodular functions

In this section we show how our new structural result can be used to provide learning guarantees in the PMAC learning model for approximate distributional learning introduced in [1]. In this model, there is a fixed but unknown distribution D over sets in $2^{[n]}$ and a fixed but unknown function $f^* : 2^{[n]} \to \mathbb{R}_+$. A learning algorithm is provided a collection $\mathcal{S} = \{S_1, S_2, \ldots\}$ of polynomially many sets drawn i.i.d. from D, as well as the value of f^* at each $S_i \in \mathcal{S}$. The goal is to design a polynomial-time algorithm that outputs a function f such that, with large probability over $\mathcal{S}$, the set of sets for which f is a good approximation for f^* has large measure with respect to D; the function f should also be evaluatable in polynomial time. More formally, the approximation guarantee is

$$\Pr_{S_1,S_2,\ldots\sim D}\Big[\ \Pr_{S\sim D}\left[\, f(S) \leq f^*(S) \leq \alpha f(S) \,\right] \geq 1-\epsilon \ \Big] \geq 1-\delta,$$

where f is the output of the learning algorithm when given inputs $\{\, (S_i, f^*(S_i)) \,\}_{i=1,2,\ldots}$. The approximation factor $\alpha \geq 1$ allows for multiplicative error in the function values.

Recall that the traditional PAC model requires one to predict the value exactly on a set of large measure and with high confidence. In contrast, the PMAC model requires one to approximate the value of a function on a set of large measure and with high confidence. Asking for low multiplicative error on most points composes naturally with approximation algorithm guarantees.

Theorem 3. Let $\mathcal{F}$ be the class of symmetric, submodular functions functions over $X = 2^{[n]}$ that take non-zero values on all sets except $\emptyset$ and $[n]$. There is an algorithm that PMAC-learns $\mathcal{F}$ with approximation factor $\sqrt{n+1}$. That is, for any distribution D over X, for any ϵ, δ sufficiently small, with probability $1-\delta$, the algorithm produces a function f that approximates f^* within a multiplicative factor of $\sqrt{n+1}$ on a set of measure $1-\epsilon$ with respect to D. The algorithm uses $m = O\left(\frac{n}{\epsilon}\log\left(\frac{n}{\delta\epsilon}\right)\right)$ training examples and runs in time $\mathrm{poly}(n, 1/\epsilon, 1/\delta)$.

Proof Sketch. We use Algorithm 1 to PMAC-learn such functions to with approximation factor $\sqrt{n+1}$.

Let $f : 2^{[n]} \to \mathbb{R}_+$ be the symmetric submodular target function. By Theorem 2 we know that there exists a function $\hat{f}$ of the form $\hat{f}(S) = \sqrt{\chi(S)^{\mathsf{T}}A\chi(S)}$ where A is a symmetric positive definite matrix such that $\hat{f}(S) \leq f(S) \leq \sqrt{n}\hat{f}(S)$ for all $S \subseteq [n]$.

ALGORITHM 1: Algorithm for PMAC-learning the class of symmetric submodular functions.

Input: A sequence of labeled training examples $\mathcal{S} = \{(S_1, f^*(S_1)), (S_2, f^*(S_2)), \ldots (S_m, f^*(S_m))\}$.

- Re-represent each training example S as $\chi_M(S)$, where is $\chi_M(S)$ is $N = \binom{n}{2} + n$-dimensional with one feature for each subset of $[n]$ with at most 2 items; for $i_1 \neq i_2$, $\chi_M(S)_{i_1,i_2} = 1$ if both items i_1 and i_2 appear in S and $\chi_M(S)_{i_1,i_2} = 0$ otherwise; also $\chi_M(S)_{i,i} = \chi(S)_i$.
- For each $1 \leq i \leq m$, let y_i be the outcome of flipping a fair $\{+1, -1\}$-valued coin, each coin flip independent of the others. Let $x_i \in \mathbb{R}^{N+1}$ be the point defined by

$$x_i = \begin{cases} (\chi_M(A_i), f^{*2}(A_i)) & (\text{if } y_i = +1) \\ (\chi_M(A_i), (n+1) \cdot f^{*2}(A_i)) & (\text{if } y_i = -1). \end{cases}$$

- Find a linear separator $u = (w, -z) \in \mathbb{R}^{N+1}$, where $w \in \mathbb{R}^N$ and $z \in \mathbb{R}$, such that u is consistent with the labeled examples (x_i, y_i) $\forall i \in [m]$.

Output: The function f defined as $f(S) = \sqrt{\frac{1}{(n+1)z} w^{\mathsf{T}} \chi_M(S)}$.

This structural result suggests re-representing each set S by a new set of $N = \binom{n}{2} + n$ features, with one feature for each subset of $[n]$ with at most 2 items. Formally, for any set $S \subseteq [n]$, we denote by $\chi_M(S)$ its feature representation over this new set of features. $\chi_M(S)_{i_1,i_2} = 1$ if both items i_1 and i_2 appear in S and $\chi_M(S)_{i_1,i_2} = 0$ otherwise. Clearly $\hat{f}(S)$ is representable as the square root of a linear function over this new set of features. Given this the argument then follows similarly to the one in [1]. In particular, note that following examples in $\mathbb{R}^{N+1}$ are linearly separable since $nw^{\mathsf{T}}\chi_M(S) - (f^*(S))^2 \geq 0$ and $nw^{\mathsf{T}}\chi_M(S) - (n+1)(f^*(S))^2 < 0$.

$$\begin{array}{lll} \text{Examples labeled } +1: & \text{ex}_S^+ := (\chi_M(S), (f^*(S))^2) & \forall S \subseteq [n] \\ \text{Examples labeled } -1: & \text{ex}_S^- := (\chi_M(S), (n+1) \cdot (f^*(S))^2) & \forall S \subseteq [n] \end{array}$$

This suggests trying to reduce our learning problem to the standard problem of learning a linear separator for these examples in the standard PAC model [6, 11]. However, in order to apply standard techniques to learn such a linear separator, we must ensure that our training examples are i.i.d. To achieve this, we create a i.i.d. distribution D' in $\mathbb{R}^{N+1}$ that is related to the original distribution D as follows. First, we draw a sample $S \subseteq [n]$ from the distribution D and then flip a fair coin for each. The sample from D' is labeled ex_S^+ i.e. $+1$ if the coin is heads and ex_S^- i.e. -1 if the coin is tails. As mentioned above, these labeled examples are linearly separable in $\mathbb{R}^{N+1}$. Conversely, suppose we can find a linear separator that classifies most of the examples coming from D' correctly. Assume that this linear separator in $\mathbb{R}^{N+1}$ is defined by the function $u^{\mathsf{T}}x = 0$, where $u = (\hat{w}, -z)$, $w \in \mathbb{R}^N$ and $z > 0$. The key observation is that the function $f(S) = \frac{1}{(n+1)z}\hat{w}^{\mathsf{T}}\chi_M(S)$ approximates $(f^*(\cdot))^2$ to within a factor $n+1$ on most of the points coming from D. ■

Interestingly, the cut function in a graph, a canonical example of a non-monotone symmetric submodular function, is a quadratic function in $\mathbb{R}^n$ since it can be written as $f(S) = \chi(S)^T L \chi(S)$, where L is the Laplacian of the graph; therefore it is PAC learnable from $\tilde{O}(n^2/\epsilon)$ examples. So it is natural to ask whether the upper bound in Theorem 3 can be improved, perhaps to obtain a constant factor approximation. We show in the following that this is not possible: symmetric submodular functions cannot be PMAC-learned with approximation factor $\tilde{o}(n^{1/3})$.

Theorem 4. The class of symmetric submodular function cannot be PMAC learned with an approximation factor $\tilde{o}(n^{1/3})$.

Proof. We provide an approximation-preserving reduction from the problem of PMAC learning the class of monotone submodular functions to the problem of PMAC learning the class of symmetric non-monotone submodular functions. Because of the PMAC learning lower bound of Balcan and Harvey [1], which showed that monotone, submodular functions cannot be PMAC-learned with approximation factor $\tilde{o}(n^{1/3})$, the reduction implies the desired result.

The key in our reduction is showing how for any monotone submodular function f there is a related submodular function g such that knowing the values of one of them completely determines the values of the other. [1] For a set S and integer i, let $S + i$ denote $S \cup \{i\}$.

[1] We thank M. Queyranne for pointing this classic claim.

Claim 5. Let $f : 2^{[n]} \to \mathbb{R}$ be a monotone, submodular function. Define $g : 2^{[n]+0} \to \mathbb{R}$ by $g(S) = f(S)$ and $g(S+0) = f([n] \setminus S)$ for all $S \subseteq \{1, \ldots, n\}$. Then g is symmetric and submodular.

To complete the proof of Theorem 4 it suffices to prove the claim. Clearly g is symmetric, so we check submodularity.

Case 1: $0 \notin A \cup B$. This follows from submodularity of f.

Case 2: $0 \in A \cap B$. This follows from submodularity of the map $S \mapsto f([n] \setminus S)$.

Case 3: Lastly we consider the case that only one set contains 0. For $A, B \subseteq [n]$, we have

$$\begin{aligned} g(A) + g(B+0) &= f(A) + f([n] \setminus B) \geq f(A \cap B) + f\big(([n] \setminus A) \cap ([n] \setminus B)\big) \\ &= g(A \cap B) + g\big((A \cup B) + 0\big) = g\big(A \cap (B+0)\big) + g\big(A \cup (B+0)\big), \end{aligned}$$

where the inequality is by monotonicity of f. ■

5 Discussion and Open Problems

The most natural open question raised by our work is whether there an algorithm for learning arbitrary non-monotone submodular functions in the PMAC model to within an approximation factor $O(n^{1/2})$. Another interesting open question is whether symmetric submodular functions can be approximately learned everywhere with value queries with a factor of $\tilde{O}(n^{1/2})$ (Note that by using a reduction similar to the one in Theorem 4 and a result of [4] we can show a lower bound of $\tilde{o}(n^{1/2})$ for learning symmetric submodular functions in this model). For the simpler case of monotone submodular functions, Goemans et al. [4] use their structural result for monotone submodular functions to give an algorithm for approximately learning everywhere with value queries. To do so, Goemans et al. compute an approximation to the John ellipsoid E for a convex body derived from EP_f, given an oracle for approximately maximizing certain norms over that body. Their task is made simpler by the fact that their convex body has an axis-aligned John ellipsoid. For symmetric submodular functions this is more challenging because the John ellipsoid of the polytope B_f need not be axis-aligned. It would be very interesting to resolve whether there is an efficient algorithm to compute a poly-logarithmic approximation to the John ellipsoid for B_f, where f is a symmetric, submodular function. Finally, it would be interesting to close the gap between the $O(n^{1/2})$ upper bound and our $\tilde{o}(n^{1/3})$ of our lower bound in the PMAC model .

References

[1] M.-F. Balcan and N. Harvey. Learning submodular functions. In *STOC*, 2011.

[2] M. Cheraghchi, A. R. Klivans, P. Kothari, and H. K. Lee. Submodular functions are noise stable. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2012.

[3] S. Fujishige. Canonical decompositions of symmetric submodular functions. *Discrete Applied Mathematics*, 5:175–190, 1983.

[4] M. Goemans, N. Harvey, S. Iwata, and V. Mirrokni. Approximating submodular functions everywhere. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2009.

[5] A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately releasing conjunctions and the statistical query barrier. In *STOC*, 2011.

[6] M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. 1994.

[7] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, 2005.

[8] M. Narasimhan and J. Bilmes. Local search for balanced submodular clusterings. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 2007.

[9] M. Queyranne. Minimizing symmetric submodular functions. *Mathematical Programming*, 82:3–12, 1998.

[10] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer, 2004.

[11] V. N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.