Approximately Adaptive Submodular Maximization

Matt J. Kusner
Department of Computer Science & Engineering
Washington University in St. Louis
St. Louis, MO 63130
mkusner@wustl.edu

Abstract

Recently, a powerful framework for adaptive stochastic set optimization called *adaptive submodularity* has been introduced. As a result, there has been a wealth of interest towards applying this framework to a variety of real-world problems. To do so, the modeler is tasked with discovering an adaptive submodular function that captures key aspects of the learning problem. While deriving such functions are straightforward for certain applications, demonstrating adaptive submodularity is often nontrivial. In this paper, we propose a relaxation of the adaptive submodularity framework that retains near-optimality guarantees via *approximate adaptive submodularity*. We make two main contributions: 1. we introduce the general framework of *approximately adaptive submodular* function optimization and prove a near-optimal performance guarantee for the greedy selection algorithm; 2. we show maximization of such functions under a submodular cost constraint can be performed near-optimally.

1 Introduction

Adaptive decision-making under partial observations is now a classic learning problem. Application settings are as diverse as predicting hepatitis B [18], large-scale web-search ranking [19], survival prediction for lung cancer [15], and massive email spam filters [17]. In the partial observation setting, the outcomes of unobserved states are commonly stochastic, and we assume we have (or may construct) a 'belief' about the likelihood of these outcomes. Recently, Golovin & Krause [4] derived near-optimal guarantees for optimizing a particular type of partially observable stochastic function. These functions, called *adaptive submodular set functions*, generalize a class of set functions [14] that naturally arise in a large number of real-world discrete and combinatorial problems [12; 10; 7].

This generalization (*i.e.*, extending submodular set function optimization to the adaptive setting) has been fruitful in the settings of active detection [2], viral marketing [4], and batch active learning [1] (for a larger list see [4]). However, there are a number of learning scenarios including dictionary selection [11], feature selection [3], and anytime learning [6] that admit only near-submodular set functions. Developing adaptive decision-making models that are simultaneously suitable for these learning tasks and have strong theoretical guarantees requires a framework that supports such near-submodularity.

At the same time, in many of these near-submodular settings optimization must be performed under constraints. For instance, in the feature-cost sensitive learning scenario of Kusner et al. [13] the aim is to maximize a near-submodular function under a modular constraint on the total cost of the selected features. While optimization under cardinality, modular, and matroid constraints is well-studied in adaptive submodularity [5], it is still an open problem if a near-optimal guarantee exists for the adaptive setting if the constraint is a submodular function.

In this paper we develop such a framework, which we refer to as *approximate adaptive submodularity*. Our work builds upon prior results in approximate submodularity of accuracy-related functions [3; 6] and adaptive submodularity [4]. We demonstrate that greedily selecting items to maximize the *expected marginal improvement* yields a set that is near-optimal. We make two main contributions: 1. we introduce the framework of *approximate adaptive submodularity* and prove that for this function class a greedy selection policy is near-optimal; 2. inspired by the work of Iyer & Bilmes [8] we derive guarantees for (approximate) adaptive submodular maximization, constrained by a submodular function.

2 Background

In this section we introduce our notation and provide some background on the setting of adaptive stochastic maximization. We then set up the overall optimization problem we will solve in this paper. Finally we introduce the expected marginal improvement, a crucial quantity used by the adaptive greedy techniques of Golovin & Krause [4] to yield near-optimal solutions.

Adaptive Stochastic Maximization. Our description is modeled off the explanation of Golovin & Krause [4], to which we refer the interested reader for more details. We assume we have a ground set $\mathcal{V}$ consisting of d *items*. Upon selecting an item $s_{t+1} \in \mathcal{V}$ to be in our solution set $\mathcal{S}$ we observe an associated *outcome* $\hat{x}_{s_{t+1}} \in \mathcal{X}$ (where $\mathcal{X}$ is the set of all outcomes). The overall objective in adaptive stochastic maximization is to learn a policy π for selecting a new item s_{t+1} given a set of previously-selected items $\mathcal{S}_t = \{s_1, s_2, \ldots, s_t\}$ and outcomes $\hat{\mathbf{x}}_{\mathcal{S}_t} = [\hat{x}_{s_1}, \hat{x}_{s_2}, \ldots, \hat{x}_{s_t}]^\top$. In addition, the policy π may also return the 'stop' action, which terminates the selection process. Once a new item s_{t+1} is selected by π, we are allowed to observe its outcome $\hat{x}_{s_{t+1}}$, which we can use to select subsequent items.

This policy is learned to maximize a gain function $g : 2^{\mathcal{V}\times\mathcal{X}} \rightarrow \mathbb{R}_{\geq 0}$, which takes as input a set of (item, outcome) pairs, and returns a positive real number indicating the benefit of selecting these items with certain outcomes. Let $\mathcal{S}$ denote a set of items and let their observed outcomes be $\hat{\mathbf{x}}_{\mathcal{S}}$, with a slight abuse of notation we write $g(\mathcal{S}, \hat{\mathbf{x}}_{\mathcal{S}}) \triangleq g(\{(v_i, \hat{\mathbf{x}}(v_i)) : v_i \in \mathcal{S}\})$.

Given a distribution, or belief, over the outcomes $\hat{\mathbf{x}} \in \mathcal{X}^d$ of all items in $\mathcal{V}$: $p(\hat{\mathbf{x}})$, the expected gain $\bar{g}(\cdot)$ of a policy π is as follows: $\bar{g}(\pi) \triangleq \mathbb{E}[g(\mathcal{S}_{\pi,\hat{\mathbf{x}}}, \hat{\mathbf{x}})]_{p(\hat{\mathbf{x}})}$ where $\mathcal{S}_{\pi,\hat{\mathbf{x}}}$ is the set of items selected by repeatedly applying policy π starting from an empty set of observations until the 'stop' action is reached, given a possible set of outcomes $\hat{\mathbf{x}}$.

In the modular constrained setting we can think of each item $s \in \mathcal{V}$ as having a non-negative cost c_s. The cost of an arbitrary set $\mathcal{S}$ is given by the function $c(\mathcal{S}) = \sum_{s\in\mathcal{S}} c_s$. Our goal then is to find a policy π that maximizes the expected gain and ensures that the total cost of the items selected by π stays within a pre-defined budget $B \geq 0$. Formally, our optimization becomes,

$$\max_{\pi\in\Pi} \bar{g}(\pi) \;\; \text{subject to,} \;\; \bar{c}(\mathcal{S}_{\pi,\hat{\mathbf{x}}}) \leq B, \forall \hat{\mathbf{x}}, \tag{1}$$

where Π is the set of all possible adaptive policies. To avoid always undershooting the budget B, Golovin & Krause [4] relax the budget slightly and, if the last item s_l we pick would go over budget (given already-selected items S_t), select it with probability $\tau/c(s_l)$ where $\tau = B - c(S_t)$. Then the expected cost of the policy π (for any $\hat{\mathbf{x}}$) is $\bar{c}(\mathcal{S}_{\pi,\hat{\mathbf{x}}}) = c(S_t) + (\tau/c(s_l))c(s_l) = B$. Finding the optimal policy is a combinatorial optimization problem that in the general case is very difficult to approximate [4]. However, a greedy policy can solve (1) up to an approximation, if $g(\cdot,\cdot)$ has certain desirable (albeit potentially strict) properties.

Expected Marginal Improvement [4]. Consider the following greedy policy π_G: given a set of outcomes $\hat{\mathbf{x}}_{\mathcal{S}_t}$, select a new item s_{t+1} to maximize the *expected marginal improvement* in $g(\cdot,\cdot)$ per cost. The expected marginal improvement is defined as

$$\Delta(s_{t+1} \mid \hat{\mathbf{x}}_{\mathcal{S}_t}) \triangleq \mathbb{E}\Big[g(\mathcal{S}_t \cup s_{t+1}, \hat{\mathbf{x}}_{\mathcal{S}_t\cup s_{t+1}}) - g(\mathcal{S}_t, \hat{\mathbf{x}}_{\mathcal{S}_t})\Big]_{p(\hat{\mathbf{x}}_{s_{t+1}} \mid \hat{\mathbf{x}}_{\mathcal{S}_t})}.$$

Golovin & Krause [4] show that, for a given $p(\hat{\mathbf{x}})$, if the gain g is *adaptive submodular*, *i.e.* if for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and any $e \in \mathcal{V} \setminus \mathcal{B}$ it is the case that

$$\Delta(e \mid \mathcal{B}) \leq \Delta(e \mid \mathcal{A}) \tag{2}$$

then π_G yields a $(1 - 1/e)$ approximation to the optimal policy with expected cost B. For more details see [4].

3 Approximate Adaptive Submodularity

In certain settings it may be difficult to derive an adaptive submodular gain function. In this section we extend the framework of adaptive submodularity to functions that are 'close' to adaptive submodular. We introduce near-optimal guarantees for solving the optimization problem (1) for this broader class of gain functions. Among being non-negative and generally increasing, our gain $g(\cdot,\cdot)$ need only satisfy a weak condition called ***approximate*** *adaptive submodularity*. We show that even for functions that are not strictly adaptive submodular, we can obtain strong guarantees on the adaptive greedy algorithm π_G. We start by reconsidering adaptive submodularity.

Before defining approximate adaptive submodularity, let us consider the definition of adaptive submodularity, eq. (2). This definition is a generalization of a definition for (non-adaptive) submodular set functions [14] that is identical, except it omits the expectation around the marginal gain: $g(\mathcal{B} \cup e) - g(\mathcal{B}) \leq g(\mathcal{A} \cup e) - g(\mathcal{A})$. Another way to verify that g is a (non-adaptive) submodular function is if for $\mathcal{S}, \mathcal{S}' \subseteq \mathcal{V}$ such that $\mathcal{S} \cap \mathcal{S}' = \emptyset$, it is the case that: $\sum_{s' \in \mathcal{S}'} [g(S \cup s') - g(S)] \geq g(S \cup S') - g(S)$. As such, we propose an alternative definition of an adaptive submodular function.

Definition 1. *An adaptive set function $g(\cdot,\cdot)$ with associated distribution $p(\hat{\mathbf{x}})$ is **adaptive submodular of type II** if for any two sets of observed outcomes $\hat{\mathbf{x}}_{\mathcal{S}}$ and $\hat{\mathbf{x}}_{\mathcal{S}'}$ such that selected items for each set are non-overlapping: $\mathcal{S} \cup \mathcal{S}' = \emptyset$ and $p(\hat{\mathbf{x}}_{\mathcal{S}}) > 0, p(\hat{\mathbf{x}}_{\mathcal{S}'}) > 0$, it is the case that $\sum_{s' \in \mathcal{S}'} \Delta(s' \mid \hat{\mathbf{x}}_{\mathcal{S}}) \geq \Delta(\mathcal{S}' \mid \hat{\mathbf{x}}_{\mathcal{S}})$.*

Using this definition of adaptive submodularity, inspired by ideas from Das & Kempe [3] we define a class of functions that loosely satisfy the above inequality:

Definition 2. *A function $g(\cdot,\cdot) \geq 0$ with associated distribution $p(\hat{\mathbf{x}})$ is **approximately adaptive submodular** if for any two sets of observed outcomes $\hat{\mathbf{x}}_{\mathcal{S}}$ and $\hat{\mathbf{x}}_{\mathcal{S}'}$ such that $\mathcal{S} \cap \mathcal{S}' = \emptyset$ and $p(\hat{\mathbf{x}}_{\mathcal{S}}) > 0$, $p(\hat{\mathbf{x}}_{\mathcal{S}'}) > 0$ it is the case that*

$$\sum_{s' \in \mathcal{S}'} \Delta(s' \mid \hat{\mathbf{x}}_{\mathcal{S}}) \geq \gamma \Delta(\mathcal{S}' \mid \hat{\mathbf{x}}_{\mathcal{S}}), \tag{3}$$

where $0 \leq \gamma \leq 1$ is called the submodularity ratio for non-adaptive functions [3].

The quantity γ describes how close $g(\cdot,\cdot)$, coupled with $p(\hat{\mathbf{x}})$, is to adaptive submodular. If $\gamma = 1$, $g(\cdot,\cdot)$ is fully adaptive submodular, while if $\gamma = 0$ it is nothing like an adaptive submodular function. Even though $g(\cdot,\cdot)$ is not fully submodular we introduce a novel near-optimal guarantee for the broader class of approximately adaptive submodular functions, involving the submodularity ratio,

Theorem 1. *Let $g(\cdot,\cdot)$ together with $p(\hat{\mathbf{x}})$ satisfy the following three properties,*

1. ***Adaptive Monotonicity** — for any observed outcomes $\hat{\mathbf{x}}_{\mathcal{S}_t}$ in the support of $p(\hat{\mathbf{x}})$, $g(\cdot,\cdot)$ coupled with $p(\hat{\mathbf{x}})$ is adaptive monotonic if for all $s_{t+1} \in \mathcal{V}$ it holds that, $\Delta(s_{t+1} | \hat{\mathbf{x}}_{\mathcal{S}_t}) \geq 0$.*
2. ***Approximate Adaptive Submodularity** — described above in eq. (3)*
3. ***Normalized** — if no items are selected $g(\emptyset, 0) = 0$.*

Then the greedy policy π_G that always maximizes the current marginal improvement per cost has the approximation guarantee for the optimization (1) as such

$$\bar{g}(\pi_G) > \Big(1 - e^{-\gamma}\Big)\bar{g}(\pi^*),$$

π^* is the optimal policy for (1).

We leave the proof of the theorem to the full version of the paper. Note that if $g(\cdot,\cdot)$ is adaptive submodular, we obtain $\gamma = 1$ and arrive at the same guarantee as that of adaptive submodular functions [4]. Thus this theorem generalizes the theoretical guarantees for adaptive submodular functions to adaptive stochastic maximization for 'submodular-like' adaptive functions.

4 Submodular Costs

Often, groups of items may have discounted costs, and adding an item to a group of items may cost less, as the size of the group increases. This may happen in the medical setting for example, if the results of an examination can be used as an intermediate for another test. The cost of a set of items $f : 2^{\mathcal{V}} \rightarrow \mathcal{R}_{\geq 0}$ is no longer the sum of the costs of the individual items. Instead, the increase in cost of adding an item s_{t+1} to a set of items $\mathcal{S}_t \subseteq \mathcal{V}$, may be less than adding it to a subset $\mathcal{S}_{t-i} \subseteq \mathcal{S}_t$, for $0 \leq i \leq t$. Formally, for $s_{t+1} \notin \mathcal{S}_t$

$$f(\mathcal{S}_{t-i} \cup s_{t+1}) - f(\mathcal{S}_{t-i}) \geq f(\mathcal{S}_t \cup s_{t+1}) - f(\mathcal{S}_t),$$

Thus, $f(\cdot)$ is a submodular function [14]. Further, let $f(\cdot)$ be monotonic ($f(\mathcal{S}_{t-i}) \leq f(\mathcal{S}_t)$) and normalized ($f(\emptyset) = 0$). Similar to the problem in eq. (1), our goal is to solve the following optimization

$$\max_{\pi \subseteq \Pi} \bar{g}(\pi) \;\; \text{subject to,} \;\; \bar{f}(\mathcal{S}_{\pi,\hat{\mathbf{x}}}) \leq B, \forall \hat{\mathbf{x}}, \tag{4}$$

where as before $\mathcal{S}_{\pi,\hat{\mathbf{x}}}$ is the set of items selected by π given outcomes $\hat{\mathbf{x}}$. Differently, $f(\cdot)$ is submodular, as opposed to $c(\cdot)$ which is modular. Recently, Iyer & Bilmes [8] presented an approximation guarantee for maximizing submodular functions subject to a submodular knapsack (cost) constraint. The idea is to approximate the submodular function $f(\cdot)$ with a tight modular upper bound $m^f(\cdot)$: the supergradient of $f(\cdot)$ [10]. This approximation for any set $\mathcal{S}$ is $m^f(\mathcal{S}) = \sum_{s \in \mathcal{S}} f(s)$. After this replacement, they give a guarantee for greedily selecting items to maximize the marginal improvement per cost. We extend this guarantee to approximately adaptive submodular functions.

Theorem 2. *Let a function $g(\cdot,\cdot)$ that together with distribution $p(\hat{\mathbf{x}})$ is (i) normalized, (ii) adaptive monotonic, and (iii) approximately adaptive submodular. If we replace $f(\mathcal{S})$ with the modular function $m^f(\mathcal{S}) = \sum_{s \in \mathcal{S}} f(s)$, the expected adaptive greedy policy π_G maximizing expected marginal improvement per (modular) cost approximates the optimal expected adaptive policy π^* in the following ways*

$$\bar{g}(\pi_G) > \Big(1 - e^{-\gamma}\Big)\bar{g}(\pi^*) \qquad\qquad \bar{f}(\mathcal{S}_{\pi,\hat{\mathbf{x}}}) \leq \left[\frac{d}{1 + (d-1)(1-\kappa_f)}\right] B, \forall \hat{\mathbf{x}},$$

where $d = |\mathcal{V}|$ is the set of all items and κ_f is the curvature of $f(\cdot)$ [16; 9].

Proof. In order to prove Theorem 2 we prove a slightly modified version, Lemma 1, and subsequently show how to arrive at the proof of Theorem 2. In the following proof we extend a theorem given by Iyer & Bilmes [8] to approximately adaptive submodular functions. First define $\bar{f}(\cdot) = \mathbb{E}_L[f(\cdot)]$, where the expectation is over the randomness of the policy (*i.e.,* the event L that the last item selected exceeds the budget). Recall that the modular approximation $m^f(\cdot)$ of $f(\cdot)$ is $m^f(\mathcal{S}) = \sum_{s \in \mathcal{S}} f(s)$.

Lemma 1. ***(based on Iyer & Bilmes [8], Theorem 4.8)*** *Replacing $f(\cdot)$ with $m^f(\cdot)$ and running the greedy policy π with modular cost constraint for all $\hat{\mathbf{x}}$: $\mathbb{E}_L[m^f(\mathcal{S}_{\pi,\hat{\mathbf{x}}})] = m^{\bar{f}}(\mathcal{S}_{\pi,\hat{\mathbf{x}}}) \leq B$, gives an approximation $\bar{g}(\pi) > (1 - e^{-\gamma})\bar{g}(\tilde{\pi})$, where $\tilde{\pi}$ is the optimal solution of*

$$\max_{\pi' \in \Pi} \bar{g}(\pi') \;\; \textit{subject to,} \;\; \bar{f}(\mathcal{S}_{\pi',\hat{\mathbf{x}}}) \leq B\left[\frac{1 + (d-1)(1-\kappa_f)}{d}\right], \; \forall \hat{\mathbf{x}} \tag{5}$$

where $d = |\mathcal{V}|$ and κ_f is the curvature of $f(\cdot)$ [16; 9].

Proof of Lemma 1. We are given that $\tilde{\pi}$ is the optimal policy for the optimization in (5). Notice that $\tilde{\pi}$ is also a feasible policy for the optimization,

$$\max_{\pi' \in \Pi} \bar{g}(\pi') \;\; \text{subject to,} \;\; \mathbb{E}_L\left[\sum_{s'_i \in \mathcal{S}_{\pi',\hat{\mathbf{x}}}} f(s'_i)\right] = m^{\bar{f}}(\mathcal{S}_{\pi',\hat{\mathbf{x}}}) \leq B, \; \forall \hat{\mathbf{x}}$$

where s'_i denotes the ith element selected by the policy π'. The reason $\tilde{\pi}$ is also a feasible policy for the above optimization is because for $\tilde{\pi}$,

$$\begin{aligned}
\mathbb{E}_L\left[\sum_{\tilde{s}_i \in \mathcal{S}_{\tilde{\pi},\hat{\mathbf{x}}}} f(\tilde{s}_i)\right] &\leq \mathbb{E}_L\left[\frac{K_{\bar{f},\hat{\mathbf{x}}}}{1 + (K_{\bar{f},\hat{\mathbf{x}}} - 1)(1-\kappa_f)} f(\mathcal{S}_{\tilde{\pi},\hat{\mathbf{x}}})\right] \\
&\leq \frac{d}{1 + (d-1)(1-\kappa_f)} \mathbb{E}_L\Big[f(\mathcal{S}_{\tilde{\pi},\hat{\mathbf{x}}})\Big] = \frac{d}{1 + (d-1)(1-\kappa_f)} \bar{f}(\mathcal{S}_{\tilde{\pi},\hat{\mathbf{x}}}) \\
&\leq \frac{d}{1 + (d-1)(1-\kappa_f)} \left[\frac{1 + (d-1)(1-\kappa_f)}{d}\right] B = B,
\end{aligned}$$

where $K_{\bar{f},\hat{\mathbf{x}}} = \max\{|\mathcal{S}_{\pi',\hat{\mathbf{x}}}| : \bar{f}(\mathcal{S}_{\pi',\hat{\mathbf{x}}}) \leq B\}$. The first inequality is a general bound on the modular approximation of a submodular function $f(\cdot)$ given by Iyer et al. [9]. The second line follows as the largest possible value of $K_{\bar{f},\hat{\mathbf{x}}}$ is d and the function $\frac{x}{1+(x-1)(1-\kappa_f)}$ is monotonically increasing in x for $x \geq 0$ given that $0 \leq \kappa_f \leq 1$ [8], and from the definition of $\bar{f}(\cdot)$. The third line is given by the definition of $\tilde{\pi}$ as an optimizer of (5).

Now, using greedy selection with $m^{\bar{f}}(\cdot)$ replacing $\bar{f}(\cdot)$ we are guaranteed to find a policy π such that $\bar{g}(\pi) > (1 - e^{-\gamma})\bar{g}(\tilde{\pi})$ with constraint $m^{\bar{f}}(\mathcal{S}_{\pi,\hat{\mathbf{x}}}) \leq B, \forall\hat{\mathbf{x}}$ as given by Theorem 1. □

Proof of Theorem 2 (continued). To finally prove Theorem 2 we note that if we instead run the greedy policy π to solve the following optimization,

$$\max_{\pi' \in \Pi} \bar{g}(\pi') \text{ subject to, } m^{\bar{f}}(\mathcal{S}_{\pi',\hat{\mathbf{x}}}) \leq \left[\frac{d}{1 + (d-1)(1-\kappa_f)}\right] B, \forall\hat{\mathbf{x}}$$

where we have replaced budget B with new budget $B' = \left[\frac{d}{1+(d-1)(1-\kappa_f)}\right]B$, then we arrive at an approximation $\bar{g}(\pi) > (1 - e^{-\gamma})\bar{g}(\tilde{\pi})$, where $\tilde{\pi}$ is the optimal solution of

$$\max_{\pi' \in \Pi} \bar{g}(\pi') \text{ subject to, } \bar{f}(\mathcal{S}_{\pi',\hat{\mathbf{x}}}) \leq B' \left[\frac{1 + (d-1)(1-\kappa_g)}{d}\right] = B, \forall\hat{\mathbf{x}}$$

as the fractions cancel out. Therefore, $\pi^* = \tilde{\pi}$ and we are guaranteed that $\bar{f}(\mathcal{S}_{\pi',\hat{\mathbf{x}}}) \leq \left[\frac{d}{1+(d-1)(1-\kappa_f)}\right]B, \forall\hat{\mathbf{x}}$, as $m^{\bar{f}}(\cdot)$ is a supergradient of $\bar{f}(\cdot)$. □

5 Conclusion

There are a number of directions for future work, including proving or disproving that approximate adaptive submodular guarantees can be derived for the batch selection strategy [1] (*i.e.,* such that outcomes are only observed after an entire batch of items is selected). Additionally, it is yet unknown if it is possible to derive similar guarantees for approximate adaptive submodular minimization. To summarize, we have introduced a relaxation of the framework of adaptive submodularity for functions that are 'close' to such functions. This relaxation is natural for problems in feature selection and anytime learning. Overall, it is our hope that practitioners and researchers alike will find this framework a ready tool for deriving new near-optimal adaptive models.

References

[1] Chen, Y. and Krause, A. Near-optimal batch mode active learning and adaptive submodular optimization. In *ICML*, 2013.

[2] Chen, Yuxin, Shioi, Hiroaki, Montesinos, Cesar Fuentes, Koh, Lian Pin, Wich, Serge, and Krause, Andreas. Active detection via adaptive submodularity. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 55–63, 2014.

[3] Das, A. and Kempe, D. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *ICML*, 2011.

[4] Golovin, D. and Krause, A. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *JAIR*, 2011.

[5] Golovin, Daniel and Krause, Andreas. Adaptive submodular optimization under matroid constraints. *arXiv preprint arXiv:1101.4450*, 2011.

[6] Grubb, A. and Bagnell, J. A. Speedboost: Anytime prediction with uniform near-optimality. In *AISTATS*, 2012.

[7] Iyer, Rishabh and Bilmes, Jeff. The lovász-bregman divergence and connections to rank aggregation, clustering and web ranking. In *Uncertainty in Artificial Intelligence (UAI)*, Bellevue, Washington, July 2013. AUAI.

[8] Iyer, R.K. and Bilmes, J.A. Submodular optimization with submodular cover and submodular knapsack constraints. In *NIPS*, 2013.

[9] Iyer, R.K., Jegelka, S., and Bilmes, J.A. Curvature and optimal algorithms for learning and minimizing submodular functions. In *NIPS*, 2013.

[10] Jegelka, Stefanie and Bilmes, Jeff. Submodularity beyond submodular energies: coupling edges in graph cuts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1897–1904. IEEE, 2011.

[11] Krause, Andreas and Cevher, Volkan. Submodular dictionary selection for sparse representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 567–574, 2010.

[12] Krause, Andreas, Singh, Ajit, and Guestrin, Carlos. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

[13] Kusner, Matt J, Chen, Wenlin, Zhou, Quan, Xu, Zhixiang Eddie, Weinberger, Kilian Q, and Chen, Yixin. Feature-cost sensitive learning with submodular trees of classifiers. In *AAAI*, 2014.

[14] Nemhauser, G.L., Wolsey, L.A., and Fisher, M.L. An analysis of approximations for maximizing submodular set functions i. *Mathematical Programming*, 1978.

[15] Raykar, V.C., Krishnapuram, B., and Yu, S. Designing efficient cascaded classifiers: tradeoff between accuracy and cost. In *SIGKDD*, pp. 853–860, 2010.

[16] Vondrák, J. Submodularity and curvature: the optimal algorithm. *RIMS Kokyuroku Bessatsu B*, 2010.

[17] Weinberger, K.Q., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. Feature hashing for large scale multitask learning. In *ICML*, pp. 1113–1120, 2009.

[18] Ye, Q.H., Qin, L.X., Forgues, M., He, P., Kim, J.W., Peng, A.C., Simon, R., Li, Y., Robles, A.I., Chen, Y., et al. Predicting hepatitis b virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nature medicine*, 9(4):416–423, 2003.

[19] Zheng, Z., Zha, H., Zhang, T., Chapelle, O., Chen, K., and Sun, G. A general boosting method and its application to learning ranking functions for web search. In *NIPS*, pp. 1697–1704. Cambridge, MA, 2008.