Distributionally Robust Submodular Maximization

Matthew Staib* MIT CSAIL

Bryan Wilder* USC

Stefanie Jegelka MIT CSAIL

Abstract

Submodular functions have applications throughout machine learning, but in many settings, we do not have *direct* access to the underlying function f. We focus on *stochastic* functions that are given as an expectation of functions over a distribution P. In practice, we often have only a limited set of samples f_i from P. The standard approach *indirectly* optimizes f by maximizing the sum of f_i. However, this ignores generalization to the true (unknown) distribution. In this paper, we achieve better performance on the actual underlying function f by directly optimizing a combination of bias and variance. Algorithmically, we accomplish this by showing how to carry out distributionally robust optimization (DRO) for submodular functions, providing efficient algorithms backed by theoretical guarantees which leverage several novel contributions to the general theory of DRO. We also show compelling empirical evidence that DRO improves generalization to the unknown stochastic submodular function.

1 Introduction

Submodular functions have natural applications in many facets of machine learning and related areas, e.g. dictionary learning (Das & Kempe, 2011), influence maximization (Kempe et al., 2003; Domingos & Richardson, 2001), data summarization (Lin & Bilmes, 2011), probabilistic modeling (Djolonga & Krause, 2014) and diversity (Kulesza & Taskar, 2012). In these settings, we have a set function $f(S)$ over subsets S of some ground set of items V, and seek S^* so that $f(S^*)$ is as large or small as possible. While optimization of set functions is hard in general, submodularity enables exact minimization and approximate maximization in polynomial time.

In many settings, the submodular function we wish to optimize has additional structure, which may present both challenges and an opportunity to do better. In particular, the stochastic case has recently gained attention, where we wish to optimize $f_P(S) := \mathbb{E}_{f\sim P}[f(S)]$ for some distribution P. The most naive approach is to draw many samples from P and optimize their average; this is guaranteed to work when the number of samples is very large. Much recent work has focused on more computationally efficient gradient-based algorithms for stochastic submodular optimization (Karimi et al., 2017; Mokhtari et al., 2018; Hassani et al., 2017). All of this work assumes that we have access to a sampling oracle for P that, on demand, generates as many iid samples as are required. But in many realistic settings, this assumption fails: we may only have access to historical data and not a simulator for the ground truth distribution. Or, computational limitations may prevent drawing many samples if P is expensive to simulate.

Here, we address this gap and consider the maximization of a stochastic submodular function given access to a *fixed* set of samples $f_1, \dots, f_n$ that form an empirical distribution $\hat{P}_n$. This setup introduces elements of statistical learning into the optimization. Specifically, we need to ensure that the solution we choose generalizes well to the unknown distribution P. A natural approach is to optimize the empirical estimate $\hat{f}_n = \frac{1}{n}\sum_{i=1}^n f_i$, analogous to empirical risk minimization. The average $\hat{f}_n$ is an unbiased estimator of f_P, and when n is very large, generalization is guaranteed by standard concentration bounds. We ask: is it possible to do better, particularly in the realistic case where n is small (at least relative to the variance of P)? In this regime, a biased estimator could achieve much lower variance and thereby improve optimization.

Optimizing this bias-variance tradeoff is at the heart of statistical learning. Concretely, instead of optimizing the finite sum, we will optimize the variance-

*Equal contribution. Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

regularized objective $f_{\hat{P}_n}(S) - C_1\sqrt{\mathrm{Var}_{\hat{P}_n}(f(S))/n}$. When the variance is high, this term dominates a standard high-probability lower bound on $f_P(S)$. Unfortunately, direct optimization of this bound is in general intractable: even if all f_i are submodular, their variance need not be (Staib & Jegelka, 2017).

In the continuous setting, it is known that variance regularization is equivalent to solving a distributionally robust problem, where an adversary perturbs the empirical sample within a small ball (Gotoh et al., 2015; Lam, 2016; Namkoong & Duchi, 2017). The resulting maximin problem is particularly nice in the concave case, since the pointwise minimum of concave functions is still concave and hence global optimization remains tractable. However, this property does not hold for submodular functions, prompting much recent work on robust submodular optimization (Krause et al., 2011; Chen et al., 2017; Staib & Jegelka, 2017; Anari et al., 2017; Wilder, 2018; Orlin et al., 2016; Bogunovic et al., 2017; Mirzasoleiman et al., 2017; Kazemi et al., 2018).

In this work, **1.** we show that, perhaps surprisingly, variance-regularized submodular maximization is both tractable and scalable. **2.** We give a theoretically-backed algorithm for distributionally robust submodular optimization which substantially improves over a naive application of previous approaches for robust submodular problems. Along the way, **3.** we develop improved technical results for general (non-submodular) distributionally robust optimization problems, including both improved algorithmic tools and more refined structural characterizations of the problem. For instance, we give a more complete characterization of the relationship between distributional robustness and variance regularization. **4.** We verify empirically that in many real-world settings, variance regularization enables better generalization from fixed samples of a stochastic submodular function, particularly when the variance is high.

Related Work. We build on and significantly extend a recent line of research in statistical learning and optimization that develops a relationship between distributional robustness and variance-based regularization (Maurer & Pontil, 2009; Gotoh et al., 2015; Lam, 2016; Duchi et al., 2016; Namkoong & Duchi, 2017). While previous work has uniformly focused on the continuous (and typically convex) case, here we address *combinatorial* problems with submodular structure, requiring further technical developments. As a byproduct, we better characterize the behavior of the DRO problem under low sample variance (which was left open in previous work), show conditions under which the DRO problem becomes smooth, and provide improved algorithmic tools which apply to general DRO problems.

Another related area is robust submodular optimization (Krause et al., 2011; Chen et al., 2017; Staib & Jegelka, 2017; Anari et al., 2017; Wilder, 2018; Orlin et al., 2016; Bogunovic et al., 2017). Much of this recent surge in interest is inspired by applications to robust influence maximization (Chen et al., 2016; He & Kempe, 2016; Lowalekar et al., 2016). Existing work aims to maximize the minimum of a set of submodular functions, but does not address the *distributionally* robust optimization problem where an adversary perturbs the empirical distribution. We develop scalable algorithms, accompanied by approximation guarantees, for this case. Our algorithms improve both theoretically and empirically over naive application of previous robust submodular optimization algorithms to DRO. Further, our work is motivated by the connection between distributional robustness and generalization in learning, which has not previously been studied for submodular functions. Stan et al. (2017) study generalization in a related combinatorial problem, but they do not explicitly balance bias and variance, and the goal is different: they seek a smaller ground set which still contains a good subset for each user in the population.

A complementary line of work concerns *stochastic* submodular optimization (Mokhtari et al., 2018; Hassani et al., 2017; Karimi et al., 2017) that, as opposed to our setting, requires a sampling oracle for the underlying function. We draw from stochastic optimization tools, but assume only a fixed dataset is available.

Our motivation also relates to optimization from samples, where we have access to values of a fixed unknown function on inputs sampled from a distribution. Balkanski et al. (2017, 2016) prove hardness results for general submodular maximization from samples, with positive results for functions with bounded curvature. We address a different model where the underlying function itself is stochastic and we observe realizations of it. Hence, it is possible to well-approximate the optimization problem from polynomially many samples. The challenge is to construct algorithms that make more effective use of data.

2 Stochastic Submodular Functions and Distributional Robustness

A set function $f : 2^V \to \mathbb{R}$ is *submodular* if it satisfies *diminishing marginal gains*: for all $S \subseteq T$ and all $i \in V \backslash T$, it holds that $f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T)$. It is *monotone* if $S \subseteq T$ implies $f(S) \leq f(T)$. Let P be a distribution over monotone submodular functions

f. We assume that each function is normalized and bounded, i.e., $f(\emptyset) = 0$ and $f(S) \in [0, B]$ almost surely for all subsets S. We seek a subset S that maximizes

$$f_P(S) := \mathbb{E}_{f\sim P}[f(S)] \tag{1}$$

subject to some constraints, e.g., $|S| \leq k$. We call the function $f_P(S)$ a *stochastic submodular function*. Such functions arise in many domains; we begin with two specific motivating examples.

2.1 Stochastic Submodular Functions

Influence Maximization. Consider a graph $G = (V, E)$ on which influence propagates. We seek to choose an initial seed set $S \subseteq V$ of influenced nodes to maximize the expected number subsequently reached. Each edge can be either active, meaning that it can propagate influence, or inactive. A node is influenced if it is reachable from S via active edges. Common diffusion models specify a distribution of active edges, e.g., the Independent Cascade Model (ICM), the Linear Threshold Model (LTM), and generalizations thereof. Regardless of the specific model, each can be described by the distribution of "live-edge graphs" induced by the active edges $\mathcal{E}$ (Kempe et al., 2003). Hence, the expected number of influenced nodes $f(S)$ can be written as an expectation over live-edge graphs: $f_{\text{IM}}(S) = \mathbb{E}_{\mathcal{E}}[f(S; \mathcal{E})]$. The distribution over live-edge graphs induces a distribution P over functions f as in equation (1).

Facility Location. Fix a ground set V of possibile facility locations j. Suppose we have a (possibly infinite as in (Stan et al., 2017)) number of demand points i drawn from a distribution $\mathcal{D}$. For example, each i may correspond to a user sampled from a population $\mathcal{D}$. The goal of *facility location* is to choose a subset $S \subset V$ that covers the demand points as well as possible. Each demand point i is equipped with a vector $r^i \in \mathbb{R}^{|V|}$ describing how well point i is covered by each facility j. We wish to maximize: $f_{\text{facloc}}(S) = \mathbb{E}_{i\sim\mathcal{D}}\left[\max_{j\in S} r^i_j\right]$. Each $f(S) = \max_{j\in S} r_j$ is submodular, and $\mathcal{D}$ induces a distribution P over the functions $f(S)$ as in equation (1).

2.2 Optimization and Empirical Approximation

Two main issues arise with stochastic submodular functions. First, simple techniques such as the greedy algorithm become impractical since we must accurately compute marginal gains. Recent alternative algorithms (Karimi et al., 2017; Mokhtari et al., 2018; Hassani et al., 2017) make use of additional, specific information about the function, such as efficient gradient oracles for the multilinear extension. A second issue has so far been neglected: the degree of access we have to the underlying function (and its gradients). In many settings, we only have access to a limited, fixed number of samples, either because these samples are given as observed data or because sampling the true model is computationally prohibitive.

Formally, instead of the full distribution P, we have access to an empirical distribution $\hat{P}_n$ composed of n samples $f_1, \ldots, f_n \sim P$. One approach is to optimize

$$f_{\hat{P}_n} = \mathbb{E}_{f\sim\hat{P}_n}[f(S)] = \frac{1}{n}\sum_{i=1}^{n} f_i(S), \tag{2}$$

and hope that $f_{\hat{P}_n}$ adequately approximates f_P. This is guaranteed when n is sufficiently large. E.g., in influence maximization, for $f_{\hat{P}_n}(S)$ to approximate $f_P(S)$ within error ϵ with probability $1-\delta$, Kempe et al. (2015) show that $O\left(\frac{|V|^2}{\epsilon^2}\log\frac{1}{\delta}\right)$ samples suffice. To our knowledge, this is the tightest general bound available. Still, it easily amounts to thousands of samples even for small graphs; in many applications we would not have so much data.

The problem of maximizing $f_P(S)$ from samples greatly resembles statistical learning. Namely, if the f_i are drawn iid from P, then we can write

$$f_P(S) \geq f_{\hat{P}_n}(S) - C_1\sqrt{\frac{\text{Var}_P\left(f(S)\right)}{n}} - \frac{C_2}{n} \tag{3}$$

for each S with high probability, where C_1 and C_2 are constants that depend on the problem. For instance, if we want this bound to hold with probability $1-\delta$, then applying the Bernstein bound (see Appendix A) yields $C_1 \leq \sqrt{2\log\frac{1}{\delta}}$ and $C_2 \leq \frac{2B}{3}\log\frac{1}{\delta}$ (recall B is an upper bound on $f(S)$). Given that we have only finite samples, it would then be logical to directly optimize

$$f_{\hat{P}_n}(S) - C_1\sqrt{\text{Var}_{\hat{P}_n}\left(f(S)\right)/n}, \tag{4}$$

where $\text{Var}_{\hat{P}_n}$ refers to the empirical variance over the sample. This would allow us to directly optimize the tradeoff between bias and variance. However, even when each f is submodular, the variance-regularized objective need not be (Staib & Jegelka, 2017).

2.3 Variance regularization via distributionally robust optimization

While the optimization problem (4) is not directly solvable via submodular optimization, we will see next that distributionally robust optimization (DRO) can help provide a tractable reformulation. In DRO, we seek to optimize our function in the face of an adversary who perturbs the empirical distribution within an uncertainty set $\mathcal{P}$:

$$\max_S \min_{\tilde{P}\in\mathcal{P}} \mathbb{E}_{f\sim\tilde{P}}[f(S)]. \tag{5}$$

We focus on the case when the adversary set $\mathcal{P}$ is a χ^2 ball around the empirical distribution:

Definition 2.1. The χ^2 divergence between distributions P and Q is $D_\phi(P||Q) = \frac{1}{2}\int (dP/dQ - 1)^2\, dQ$. The χ^2 uncertainty set around an empirical distribution $\hat{P}_n$ is $\mathcal{P}_{\rho,n} = \{\tilde{P} : D_\phi(\tilde{P}||\hat{P}_n) \le \rho/n\}$. When $\hat{P}_n$ corresponds to an empirical sample $Z_1, \dots, Z_n$, we encode $\tilde{P}$ by a vector p in the simplex Δ_n and equivalently write $\mathcal{P}_{\rho,n} = \left\{p \in \Delta_n : \frac{1}{2}\|np - \mathbf{1}\|_2^2 \le \rho\right\}$.

In particular, maximizing the variance-regularized objective (4) is equivalent to solving a distributionally robust problem when the sample variance is high enough: The intuition behind this equivalence is that the χ^2 ball is a quadratic ball in the simplex, and the variance penalty is also quadratic. More formally:

Theorem 2.1 (modified from (Namkoong & Duchi, 2017)). *Fix $\rho \ge 0$, and let $Z \in [0, B]$ be a random variable (i.e. $Z = f(S)$). Write $s_n^2 = \mathrm{Var}_{\hat{P}_n}(Z)$ and let $OPT = \inf_{\tilde{P}\in\mathcal{P}_{\rho,n}} \mathbb{E}_{\tilde{P}}[Z]$. Then*

$$\max\left\{0, \sqrt{\frac{2\rho}{n}s_n^2} - \frac{2B\rho}{n}\right\} \le \mathbb{E}_{\hat{P}_n}[Z] - OPT \le \sqrt{\frac{2\rho}{n}s_n^2}.$$

Moreover, if $s_n^2 \ge 2\rho(\max_i z_i - \overline{z}_n)^2/n$, then $OPT = \mathbb{E}_{\hat{P}_n}[Z] - \sqrt{2\rho s_n^2/n}$, i.e., DRO is exactly equivalent to variance regularization.

In several settings, Namkoong & Duchi (2017) show this holds with high probability, by requiring high population variance $\mathrm{Var}_P(Z)$ and applying concentration results. While Theorem 2.1 is a direct port from the convex setting, the corresponding high probability result for submodular functions is more involved:

Lemma 2.1. *Fix parameters δ, ρ, $|V|$ and $k \ge 1$. Define the constant $M = \max\{\sqrt{32\rho/7}, \sqrt{36\left(\log(1/\delta) + |V|\log(25k)\right)}\}$. For all S with $|S| \le k$ and $\mathrm{Var}_{\mathcal{P}}(f_P(S)) \ge \frac{B}{\sqrt{n}}M$, DRO is exactly equivalent to variance regularization with combined probability at least $1 - \delta$.*

This result is obtained as a byproduct of a more general argument that applies to all points in a fractional relaxation of the submodular problem (see Appendix B) and shows equivalence of the two problems when the variance is sufficiently high. However, it is not clear what the DRO problem yields when the sample variance is too small. We give a more precise characterization of how the DRO problem behaves under arbitrary variance:

Lemma 2.2. *Let $\rho < n(n-1)/2$. Suppose all $z_1, \dots, z_n$ are distinct, with $z_1 < \cdots < z_n$. Define $\alpha(m, n, \rho) = 2\rho m/n^2 + m/n - 1$, and let $\mathcal{I} = \{m \in \{1, \dots, n\} : \alpha(m, n, \rho) > 0\}$. Then, $\inf_{\tilde{P}\in\mathcal{P}_{\rho,n}} \mathbb{E}_{\tilde{P}}[Z]$ is equal to*

$$\min_{m\in\mathcal{I}}\left\{\overline{z}_m - \min\left\{\sqrt{\alpha(m,n,\rho)s_m^2},\ \frac{s_m^2}{z_m - \overline{z}_m}\right\}\right\} \le \mathbb{E}_{\hat{P}_n}[Z] - \min\left\{\sqrt{\frac{2\rho s_n^2}{n}},\ \frac{s_n^2}{z_n - \overline{z}_n}\right\},$$

where $\hat{P}_m$ denotes the uniform distribution on $z_1, \dots, z_m$, $\overline{z}_m = \mathbb{E}_{\hat{P}_m}[Z]$, and $s_m^2 = \mathrm{Var}_{\hat{P}_m}(Z)$.

The inequality holds since n is always in $\mathcal{I}$ and $\alpha(n, n, \rho) = 2\rho/n$. As in Theorem 2.1, when the variance $s_n^2 \ge 2\rho/n \cdot (z_n - \overline{z}_n)^2$, we recover the exact variance expansion. We show Lemma 2.2 by developing an exact algorithm for optimization over the χ^2 ball (see Appendix C).

Finally, we apply the equivalence of DRO and variance regularization to obtain a surrogate optimization problem. Fix the set S, and let Z be the random variable induced by $f(S)$ with $f \sim P$. Theorem 2.1 in this setting suggests that instead of directly optimizing equation (4), we can instead solve

$$\max_S \min_{\tilde{P}\in\mathcal{P}_{\rho,n}} \mathbb{E}_{f\sim\tilde{P}}[f(S)] = \max_S \min_{p\in\mathcal{P}_{\rho,n}} \sum_{i=1}^n p_i f_i(S). \tag{6}$$

3 Algorithmic Approach

Even though each $f_i(\cdot)$ is submodular, it is not obvious how to solve Problem (6): robust submodular maximization is in general inapproximable, i.e. no polynomial-time algorithm can guarantee a positive fraction of the optimal value unless P = NP (Krause et al., 2008). Recent work has sought tractable relaxations (Staib & Jegelka, 2017; Krause et al., 2008; Wilder, 2018; Anari et al., 2017; Orlin et al., 2016; Bogunovic et al., 2017), but these either do not apply or yield much weaker results in our setting. We consider a relaxation of robust submodular maximization that returns a near-optimal *distribution* over subsets S (as in (Chen et al., 2017; Wilder, 2018)). That is, we solve the robust problem $\max_{\mathcal{D}} \min_{i\in[m]} \mathbb{E}_{S\sim\mathcal{D}}[f_i(S)]$ where $\mathcal{D}$ is a distribution over sets S. It is not immediately clear how to represent a distribution over exponentially many subsets. We will later see that optimizing a product distribution (i.e. via the multilinear extension) is enough. Our strategy, based on "continuous greedy" ideas, extends the set function f to a continuous function F, then maximizes a robust problem involving F via continuous optimization.

Multilinear extension. One canonical extension of a submodular function f to the continuous domain

is the *multilinear extension*. The multilinear extension $F : [0,1]^{|V|} \to \mathbb{R}$ of f is defined as $F(x) = \sum_{S \subseteq V} f(S) \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j)$. That is, $F(x)$ is the expected value of $f(S)$ when each item i in the ground set is included in S independently with probability x_i. A crucial property of F (that we later return to) is that it is a continuous *DR-submodular* function:

Definition 3.1. A continuous function $F : \mathcal{X} \to \mathbb{R}$ is DR-submodular if, for all $x \leq y \in \mathcal{X}$, $i \in [n]$, and $\gamma > 0$ so that $x + \gamma e_i$ and $y + \gamma e_i$ are still in $\mathcal{X}$, we have $F(x + \gamma e_i) - F(x) \geq F(y + \gamma e_i) - F(y)$.

Essentially, a DR-submodular function is concave along increasing directions. Efficient algorithms are available for maximizing DR-submodular functions over convex sets (Calinescu et al., 2011; Feldman et al., 2011; Bian et al., 2017). Specifically, we take $\mathcal{X}$ to be the convex hull of the indicator vectors of feasible sets. The robust continuous optimization problem we wish to solve is then

$$\max_{x \in \mathcal{X}} \min_{p \in \mathcal{P}_{\rho,n}} \sum_{i=1}^{n} p_i F_i(x). \tag{7}$$

It remains to address two questions: (1) how to efficiently solve Problem (7) – existing algorithms only apply to the max, not the maximin version – and (2) how to then obtain a solution for Problem (6).

We address the former question in the next section. For the latter question, given a maximizing distribution $\mathcal{D}$ over subsets, existing techniques (e.g., swap rounding) can be used to round $\mathcal{D}$ to a deterministic subset S with no loss in solution quality (Chekuri et al., 2010). But our variable x in Problem (7) can only express a limited class of distributions with independent marginals $\mathbb{P}(i \in S)$, not all distributions $\mathcal{D}$. Fortunately, this discrepancy does not cost us much:

Lemma 3.1. *Suppose x is an α-optimal solution to Problem (7). Then x induces a distribution $\mathcal{D}$ over subsets so that $\mathcal{D}$ is $(1-1/e)\alpha$-optimal for Problem (6).*

Our proof involves the *correlation gap* (Agrawal et al., 2010). It is also possible to eliminate the $(1 - 1/e)$ gap altogether by using multiple copies of the decision variables to optimize over a more expressive class of distributions (Wilder, 2018), but empirically we find this unnecessary.

Next, we address algorithms for solving Problem (7). Since a convex combination of submodular functions is still submodular, we can see each p as inducing a submodular function, so Problem (7) asks to maximize the minimum of a set of continuous submodular functions.

Frank-Wolfe algorithm and complications. In the remainder of this section, we show how Problem (7) can be solved with optimal approximation ratio (as in Lemma 3.1) by Algorithm 1, which is based on Frank-Wolfe (FW) (Frank & Wolfe, 1956; Jaggi, 2013). FW algorithms iteratively move toward the feasible point that maximizes the inner product with the gradient. Instead of a projection step, each iteration uses a single linear optimization over the feasible set $\mathcal{X}$; this is very cheap for the feasible sets we are interested in (e.g., a simple greedy algorithm for matroid polytopes). Indeed, FW is currently the best approach for maximizing DR-submodular functions in many settings. While there are FW algorithms designed for convex-concave games (Gidel et al., 2017), it is not possible to directly apply these to the submodular setting while maintaining approximation guarantees.

Algorithm 1 Momentum Frank-Wolfe (MFW) for DRO

Input: functions F_i, time T, batch size c, parameter ρ, stepsizes $\rho_t > 0$
$x^{(0)} \leftarrow \mathbf{0}$
for $t = 1, \dots, T$ **do**
  $p^{(t)} \leftarrow \operatorname{argmin}_{p \in \mathcal{P}_{\rho,n}} \sum_{i=1}^{n} p_i F_i(x^{(t-1)})$
  Draw $i_1, \dots, i_c$ from $\{1, \dots, n\}$
  $\tilde{\nabla}^{(t)} \leftarrow \frac{1}{c} \sum_{\ell=1}^{c} p_{i_\ell}^{(t)} \nabla F_{i_\ell}(x^{(t-1)})$
  $d^{(t)} \leftarrow (1 - \rho_t) d^{(t-1)} + \rho_t \tilde{\nabla}^{(t)}$
  $v^{(t)} \leftarrow \operatorname{argmax}_{v \in \mathcal{X}} \langle d^{(t)}, v \rangle$
  $x^{(t)} \leftarrow x^{(t-1)} + v^{(t)}/T$
end for
return $x^{(T)}$

Instead, observe that, since the pointwise minimum of concave functions is concave, the robust objective $G(x) = \min_{p \in \mathcal{P}_{\rho,n}} \sum_{i=1}^{n} p_i F_i(x)$ is also DR-submodular. However, a naive application of FW to $G(x)$ faces several difficulties. First, to evaluate and differentiate $G(x)$, we require an exact oracle for the inner minimization problem over p, whereas past work (Namkoong & Duchi, 2017) gave only an approximate oracle. In comparison, our submodular setting is more delicate, so an inexact oracle does not suffice: the issue is that two solutions to the inner problem can have arbitrarily *close solution value* while also providing arbitrarily *different gradients*. Hence, gradient steps with respect to an approximate minimizer may not actually improve the solution value. To resolve this issue, we provide an *exact* $O(n \log n)$ time subroutine in Appendix C. Compared to previous techniques, our algorithm rests on a more precise characterization of solutions to linear optimization over the χ^2 ball, which is often helpful in deriving structural results for general DRO problems (e.g., Lemmas 2.2 and 3.2).

Second, especially when the amount of data is large, we would like to use stochastic gradient estimates instead of requiring a full gradient computation at ev-

ery iteration. This introduces additional noise and standard Frank-Wolfe algorithms will require $O(1/\epsilon^2)$ gradient samples per iteration to cope. Accordingly, we build on a recent algorithm of Mokhtari et al. (2018) that accelerates Frank-Wolfe by re-using old gradient information; we refer to their algorithm as *Momentum Frank-Wolfe (MFW)*. For smooth DR-submodular functions, MFW achieves a $(1-1/e)$-optimal solution with additive error ϵ in $O(1/\epsilon^3)$ time. Building off MFW is advantageous versus other stochastic first-order algorithms for DR-submodular maximization, e.g. Hassani et al. (2017) achieve sub-optimal approximation ratio, and Karimi et al. (2017) focus only on a subclass of problems. Accordingly, we focus on MFW, and generalize MFW to the DRO problem by solving the next challenge.

Third, Frank-Wolfe (and MFW) require a smooth objective with Lipschitz-continuous gradients; this does *not* hold in general for pointwise minima. Wilder (2018) gets around this issue in the context of other robust submodular optimization problems by replacing $G(x)$ with the stochastically smoothed function $G_\mu(x) = \mathbb{E}_{z\sim\mu}[G(x+z)]$ as in (Duchi et al., 2012; Lan, 2013), where μ is a uniform distribution over a ball of size u. Combined with our exact inner minimization oracle, this yields a $(1-1/e)$ optimal solution to Problem (7) with ϵ error using $O(1/\epsilon^4)$ stochastic gradient samples. However, this approach results in poor empirical performance for the DRO problem (as we demonstrate later). We obtain faster convergence, in both theory and practice, through a better characterization of the DRO problem: we show that in many cases, we actually obtain a smooth problem,

Smoothness of the robust problem. While general theoretical bounds rely on smoothing $G(x)$, in practice, MFW without any smoothing performs the best. This behavior suggests that for real-world problems, the robust objective $G(x)$ may actually be smooth with Lipschitz-continuous gradient. Via our exact characterization of the worst-case distribution, we can make this intuition rigorous:

Lemma 3.2. *Define $h(z) = \min_{p\in\mathcal{P}_{\rho,n}}\langle z,p\rangle$, for $z\in[0,B]^n$, and let s_n^2 be the sample variance of z. On the subset of z's satisfying the high sample variance condition $s_n^2 \geq (2\rho B^2)/n$, $h(z)$ is smooth and has L-Lipschitz gradient with constant $L \leq \frac{2\sqrt{2\rho}}{n^{3/2}} + \frac{2}{Bn}$.*

Combined with the smoothness of each F_i, this yields smoothness of G.

Corollary 3.1. *Suppose each F_i is L_F-Lipschitz. Under the high sample variance condition, ∇G is L_G-Lipschitz for $L_G = L_F + \frac{2b\sqrt{2\rho|V|}}{n} + \frac{2b\sqrt{|V|}}{B\sqrt{n}}$.*

For submodular functions, $L_F \leq b\sqrt{k}$, where b is the largest value of a single item (Mokhtari et al., 2018). However, Corollary 3.1 is a general property of DRO (not specific to the submodular case), with broader implications. For instance, in the convex case, we immediately obtain a $O(1/\epsilon)$ convergence rate for the gradient descent algorithm proposed by Namkoong & Duchi (2017) (previously, the best possible bound would be $O(1/\epsilon^2)$ via nonsmooth techniques). Our result follows from more general properties that guarantee smoothness with fewer assumptions (see Appendices C.2, C.3). For example:

Fact 3.1. For $\rho \leq \frac{1}{2}$, the robust objective $h(z) = \min_{p\in\mathcal{P}_{\rho,n}}\langle z,p\rangle$ is smooth when $\{z_i\}$ are not all equal.

Combined with reasonable assumptions on the distribution of F_i, this means $G(x)$ is nearly always smooth. Native smoothness of the robust problem yields a significant runtime improvement over the general minimum-of-submodular case. In particular, instead of $O(1/\epsilon^4)$, we achieve the $O(1/\epsilon^3)$ rate of the simpler, non-robust submodular maximization:

Theorem 3.1. *When the high sample variance condition holds, MFW with no smoothing satisfies*

$$\mathbb{E}[G(x^{(T)})] \geq (1-1/e)\,OPT - \frac{2\sqrt{kQ}}{T^{1/3}} - \frac{Lk}{T}$$

where $Q = \max\{9^{2/3}\|\nabla G(x^0) - d^0\|, 16\sigma^2 + 3L_G^2 k\}$; σ is the variance of the stochastic gradients.

This convergence rate for DRO is almost the same as for a single submodular function (the non-robust case) (Mokhtari et al., 2018); only the Lipschitz constant is different, but this gap vanishes as n grows. It is perhaps surprising that we can obtain this rate for the robust problem, especially using an algorithm like MFW which was originally intended for the nonrobust setting. Indeed, previous work on robust submodular optimization has relied on different techniques; MFW is not an obvious candidate for DRO. However, as surveyed below, our better characterization of the DRO problem and subsequent ability to leverage MFW yields theoretical and empirical benefits.

Comparison with previous algorithms Two recently proposed algorithms for robust submodular maximization could also be used in DRO, but have drawbacks compared to MFW. Here, we compare their theoretical performance with MFW (we also show how MFW improves empirically in Section 4).

First, Chen et al. (2017) view robust optimization as a zero-sum game and apply no-regret learning to compute an approximate equilibrium. Their algorithm applies online gradient descent from the perspective of

the adversary, adjusting the distributional parameters p. At each iteration, an α-approximate oracle for submodular optimization (e.g., the greedy algorithm or a Frank-Wolfe algorithm) is used to compute a best response for the maximizing player. In order to achieve an α-approximation up to additive loss ϵ, the no-regret algorithm requires $O(1/\epsilon^2)$ iterations. However, each iteration requires a full invocation of an algorithm for submodular maximization. Our MFW algorithm requires runtime close to a *single* submodular maximization call. This results in substantially faster runtime to achieve the same solution solution quality, as we demonstrate experimentally.

Second, Wilder (2018) proposes the EQUATOR algorithm, which also applies a Frank-Wolfe approach to the multilinear extension but uses randomized smoothing. Our analysis shows smoothing is unnecessary for the DRO problem, allowing our algorithm to converge using $O(1/\epsilon^3)$ stochastic gradients, while EQUATOR requires $O(1/\epsilon^4)$. This theoretical gap is reflected in empirical performance: EQUATOR converges much more slowly, and to lower solution quality, than MFW.

4 Experiments

To probe the strength and practicality of our methods, we empirically study the two motivating problems from Section 2: influence maximization and facility location. We first report performance of distributions x^* that optimize the multilinear extension or its DRO variant (7), and later demonstrate high performance is maintained even after rounding. DRO improves test performance in all cases. All error bars are 95% confidence intervals.

4.1 Facility Location

Similar to (Mokhtari et al., 2018) we consider a facility location problem motivated by recommender systems. We use a music dataset from last.fm (las) with roughly 360000 users, 160000 bands, and over 17 million total records. For each user i, record r_j^i indicates how many times they listened to a song by band j. We seek a subset of bands so that the average user likes at least one of the bands, as measured by the playcounts. More specifically, we fix a collection of bands, and observe a *sample* of users; we seek a subset of bands that performs well on the *entire population* of users. Here, we randomly sample a subset of 1000 "train" users from the dataset, solve the DRO and ERM problems for k bands, and evaluate performance on the remaining ≈ 360000 "test" users from the dataset.

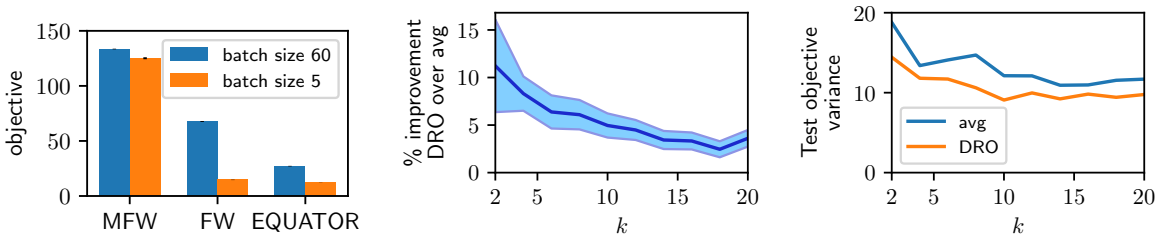
Optimization. We first compare MFW to previously proposed robust optimization algorithms, applied to the DRO problem with $k = 3$. Figure 1a compares **1.** MFW, **2.** Frank-Wolfe (FW) with no momentum and **3.** EQUATOR (Wilder, 2018). Naive FW handles noisy gradients poorly (especially with small batches), while EQUATOR underperforms since its randomized smoothing is not necessary for our natively smooth problem. We also compared to the online gradient descent (OGD) algorithm of Chen et al. (2017). OGD achieved slightly worse objective value than MFW with an order of magnitude greater runtime: OGD required 53.23 minutes on average, compared to 4.81 for MFW. EQUATOR and FW had equivalent runtime to MFW since all used the same batch size and number of iterations. MFW dominates the alternatives in both runtime and solution quality.

Generalization. Next, we evaluate the effect of DRO on test set performance across varying set sizes k. Results are averaged over 64 trials for $\rho = 10$ (corresponding to probability of failure $\delta = e^{-10}$ of the high probability bound). In Figure 1b we plot the mean percent improvement in test objective of DRO versus optimizing the average. DRO achieves clear gains, especially when the set size k is small. In Figure 1c we show the variance of test performance achieved by each method. DRO achieves lower variance, meaning that overall DRO achieves better test performance, and with better consistency.

4.2 Influence maximization

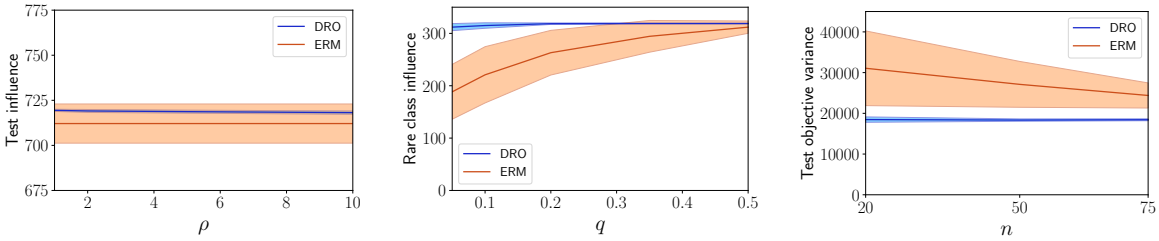
As described in Section 2, we study an influence maximization problem where we observe samples of live-edge graphs $\mathcal{E}_1, \dots, \mathcal{E}_n \sim P$. Our setting is challenging for learning: the number of samples is small and P has high variance. Specifically, P is a mixture of two different independent cascade models (ICM). In the ICM, each edge e is (independently) live with probability p_e. In our mixture, each edge has $p_e = 0.025$ with probability q and $p_e = 0.1$ with probability $1 - q$, mixing between settings of low and high influence spread. This models the realistic case where some messages are shared more widely than others. The mixture is *not* an ICM, as observing the state of one edge gives information about the propagation probability for other edges. Handling such cases is an advantage of our DRO approach over ICM-specific robust influence maximization methods (Chen et al., 2016).

We use the political blogs dataset, a network with 1490 nodes representing links between blogs related to politics (Adamic & Glance, 2005). Figure 2 compares the performance of DRO and ERM. Figure 2a shows that DRO generalizes better, achieving higher performance on the test set. Each algorithm was given $n = 20$ training samples, $k = 10$ seeds, and we set q (the frequency of low influence) to be 0.1. Test influence



(a) Algorithm comparison (b) % improvement via DRO (c) Test performance variance

Figure 1: Algorithm comparison and generalization performance on last.fm dataset.



(a) Influence on held-out set (b) Rare class (held-out) influence (c) Test performance variance

Figure 2: Influence maximization on political blogs dataset.

was evaluated via a held-out set of 3000 samples from P. Figure 2b shows that DRO's improved generalization stems from greatly improved performance on the rare class in the mixture (low propagation probabilities). For these instances, DRO obtains a greater than *40% improvement over ERM* in held-out performance for $q = 0.1$. As q increases (i.e., the rare class becomes less rare), ERM's performance on these instances converges towards DRO. A similar pattern is reflected in Figure 2c, which shows the variance in each algorithm's influence spread on the test set as a function of the number of training samples. *DRO's variance is lower by 25-40%.* As expected, DRO's advantage is greatest for small n, the most challenging setting for learning.

4.3 Rounding

Above, we report results achieved by the optimal distribution x^* on the multilinear extension $F(x^*)$ of the relevant stochastic submodular function. But to use our methods in practice, we eventually need to round x^* to a single subset S. One might worry that variabilty from the rounding procedure could erase DRO's gains. This is not the case: DRO still performs better empirically, even after rounding.

On the earlier Facility Location problem for $k = 4$, we compared the optimal distributions x^*_{ERM} and x^*_{DRO} for ERM and DRO. For each, we rounded 500 times to deterministic sets via swap rounding (Chekuri et al., 2010) and compared the resulting distributions of test objective values $\mathbb{E}_{f\sim P}[f(S)]$ (on a large subsample from the test set P). Over 64 trials (the stochasticity of MFW leads to different $x^*_{\text{ERM}}, x^*_{\text{DRO}}$), we observed that: 1. DRO always achieved better mean performance, on average by 9.3%; 2. DRO achieved lower variance in 88% of trials; 3. *for every quantile*, DRO was better on at least 73% of trials. We conclude DRO leads to better performance on the test set, both on $F(x^*)$ and on the original problem after rounding.

5 Conclusion

We address optimization of *stochastic submodular functions* $f_P(S) = \mathbb{E}_{f\sim P}[f(S)]$ in the setting where only a finite number of samples $f_1, \ldots, f_n \sim P$ is available. Instead of simply maximizing the empirical mean $\frac{1}{n}\sum_i f_i$, we directly optimize a variance-regularized version which **1.** gives a high probability lower bound for $f_P(S)$ (generalization) and **2.** allows us to trade off bias and variance in estimating f_P. We accomplish this via an equivalent reformulation as a *distributionally robust* submodular optimization problem. Along the way, we show new results for the relation between distributionally robust optimization (DRO) and variance regularization. We further give conditions for the uniqueness of the DRO solution: these are broadly useful for guaranteeing that DRO problems are smooth. Even though robust submodular maximization is hard in general, we are able to give efficient approximation algorithms for our reformulation. Empirically, our approach yields notable improvements for influence maximization and facility location problems.

Acknowledgements

This research was conducted with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a, and NSF Graduate Research Fellowship Program (GRFP). This work was supported by NSF CAREER award 1553284 and The Defense Advanced Research Projects Agency (grant number YFA17 N66001-17-1-4039). The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

References

Last.fm dataset - 360k users.

Adamic, Lada A. and Glance, Natalie. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05. 2005.

Agrawal, Shipra, Ding, Yichuan, Saberi, Amin, and Ye, Yinyu. Correlation robust stochastic optimization. In *SODA*, 2010.

Anari, Nima, Haghtalab, Nika, Naor, Joseph (Seffi), Pokutta, Sebastian, Singh, Mohit, and Torrico, Alfredo. Robust submodular maximization: Offline and online algorithms. *arXiv preprint arXiv:1710.04740*, 2017.

Balkanski, Eric, Rubinstein, Aviad, and Singer, Yaron. The power of optimization from samples. In *Advances In Neural Information Processing Systems*, 2016.

Balkanski, Eric, Rubinstein, Aviad, and Singer, Yaron. The limitations of optimization from samples. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2017.

Bian, Andrew An, Mirzasoleiman, Baharan, Buhmann, Joachim M., and Krause, Andreas. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *AISTATS*, 2017.

Bogunovic, Ilija, Mitrović, Slobodan, Scarlett, Jonathan, and Cevher, Volkan. Robust submodular maximization: A non-uniform partitioning approach. In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. 2017.

Calinescu, Gruia, Chekuri, Chandra, Pál, Martin, and Vondrák, Jan. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

Chekuri, C., Vondrak, J., and Zenklusen, R. Dependent randomized rounding via exchange properties of combinatorial structures. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 2010.

Chen, Robert S, Lucier, Brendan, Singer, Yaron, and Syrgkanis, Vasilis. Robust optimization for non-convex objectives. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

Chen, Wei, Lin, Tian, Tan, Zihan, Zhao, Mingfei, and Zhou, Xuren. Robust influence maximization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

Das, Abhimanyu and Kempe, David. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In Getoor, Lise and Scheffer, Tobias (eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11. 2011.

Djolonga, Josip and Krause, Andreas. From map to marginals: Variational inference in bayesian submodular models. In *Advances in Neural Information Processing Systems*, 2014.

Domingos, Pedro and Richardson, Matt. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001.

Duchi, John, Shalev-Shwartz, Shai, Singer, Yoram, and Chandra, Tushar. Efficient projections onto the l 1-ball for learning in high dimensions. 2008.

Duchi, John, Glynn, Peter, and Namkoong, Hongseok. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

Duchi, John C, Bartlett, Peter L, and Wainwright, Martin J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2): 674–701, 2012.

Feldman, Moran, Naor, Joseph (Seffi), and Schwartz, Roy. A unified continuous greedy algorithm for submodular maximization. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2011.

Frank, Marguerite and Wolfe, Philip. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Gidel, Gauthier, Jebara, Tony, and Lacoste-Julien, Simon. Frank-Wolfe Algorithms for Saddle Point Problems. In Singh, Aarti and Zhu, Jerry (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*. 2017.

Gotoh, Jun-ya, Kim, Michael, and Lim, Andrew. Robust Empirical Optimization is Almost the Same As Mean-Variance Optimization. *Available at SSRN 2827400*, 2015.

Hassani, Hamed, Soltanolkotabi, Mahdi, and Karbasi, Amin. Gradient Methods for Submodular Maximization. In *Advances in Neural Information Processing Systems 30*, 2017.

He, Xinran and Kempe, David. Robust influence maximization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

Jaggi, Martin. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. 2013.

Karimi, Mohammad, Lucic, Mario, Hassani, Hamed, and Krause, Andreas. Stochastic Submodular Maximization: The Case of Coverage Functions. In *Advances in Neural Information Processing Systems 30*, 2017.

Kazemi, Ehsan, Zadimoghaddam, Morteza, and Karbasi, Amin. Scalable deletion-robust submodular maximization: Data summarization with privacy and fairness constraints. In Dy, Jennifer and Krause, Andreas (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. 2018.

Kempe, David, Kleinberg, Jon, and Tardos, Éva. Maximizing the Spread of Influence Through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03. 2003.

Kempe, David, Kleinberg, Jon M, and Tardos, Éva. Maximizing the spread of influence through a social network. *Theory of Computing*, 11(4):105–147, 2015.

Krause, Andreas, McMahan, H Brendan, Guestrin, Carlos, and Gupta, Anupam. Robust submodular observation selection. *Journal of Machine Learning Research*, 9(Dec):2761–2801, 2008.

Krause, Andreas, Roper, Alex, and Golovin, Daniel. Randomized sensing in adversarial environments. In *IJCAI*, 2011.

Kulesza, Alex and Taskar, Ben. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., 2012.

Lam, Henry. Robust Sensitivity Analysis for Stochastic Systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

Lan, Guanghui. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

Lin, Hui and Bilmes, Jeff. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11. 2011.

Lowalekar, Meghna, Varakantham, Pradeep, and Kumar, Akshat. Robust Influence Maximization: (Extended Abstract). In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16. 2016.

Maurer, Andreas and Pontil, Massimiliano. Empirical bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

Mirzasoleiman, Baharan, Karbasi, Amin, and Krause, Andreas. Deletion-robust submodular maximization: Data summarization with "the right to be forgotten". In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. 2017.

Mokhtari, Aryan, Hassani, Hamed, and Karbasi, Amin. Conditional gradient method for stochastic submodular maximization: Closing the gap. In Storkey, Amos and Perez-Cruz, Fernando (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*. 2018.

Namkoong, Hongseok and Duchi, John C. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, 2016.

Namkoong, Hongseok and Duchi, John C. Variance-based Regularization with Convex Objectives. In *Advances in Neural Information Processing Systems 30*, 2017.

Orlin, James B., Schulz, Andreas, and Udwani, Rajan. Robust monotone submodular function maximization. In *Conference on Integer Programming and Combinatorial Optimization (IPCO)*, 2016.

Staib, Matthew and Jegelka, Stefanie. Robust budget allocation via continuous submodular functions. In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. 2017.

Stan, Serban, Zadimoghaddam, Morteza, Krause, Andreas, and Karbasi, Amin. Probabilistic submodular maximization in sub-linear time. In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. 2017.

Wainwright, Martin. *High-dimensional statistics: A non-asymptotic viewpoint*. 2017.

Wilder, Bryan. Equilibrium computation and robust optimization in zero sum games with submodular

structure. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.