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 $i \in V \setminus T$, 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: $f_{\mathrm{IM}}(S) = \mathbb{E}_{\mathcal{E}}[f(S; \mathcal{E})]$. 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: $f_{\mathrm{facloc}}(S) = \mathbb{E}_{i \sim \mathcal{D}}\left[\max_{j \in S} r^i_j\right]$. 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

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

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

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

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

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

where $\mathrm{Var}_{\hat{P}_n}$ 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 $\mathcal{P}_{\rho,n} = \{\tilde{P} : D_\phi(\tilde{P}||\hat{P}_n) \leq \rho/n\}$. When $\hat{P}_n$ corresponds to an empirical sample $Z_1, \ldots, Z_n$, we encode $\tilde{P}$ by a vector p in the simplex Δ_n and equivalently write $\mathcal{P}_{\rho,n} = \left\{p \in \Delta_n : \frac{1}{2}\|np - \mathbf{1}\|_2^2 \leq \rho\right\}$.

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 $\rho \geq 0$, 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*

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

Moreover, if $s_n^2 \geq 2\rho(\max_i z_i - \overline{z}_n)^2/n$, 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 $k \geq 1$. Define the constant $M = \max\{\sqrt{32\rho/7}, \sqrt{36\,(\log(1/\delta) + |V|\log(25k))}\}$. For all S with $|S| \leq k$ and $\mathrm{Var}_{\mathcal{P}}(f_P(S)) \geq \frac{B}{\sqrt{n}} M$, 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 $z_1, \ldots, z_n$ are distinct, with $z_1 < \cdots < z_n$. Define $\alpha(m, n, \rho) = 2\rho m/n^2 + m/n - 1$, and let $\mathcal{I} = \{m \in \{1, \ldots, n\} : \alpha(m, n, \rho) > 0\}$. Then, $\inf_{\tilde{P} \in \mathcal{P}_{\rho,n}} \mathbb{E}_{\tilde{P}}[Z]$ is equal to*

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

where $\hat{P}_m$ denotes the uniform distribution on $z_1, \ldots, z_m$, $\overline{z}_m = \mathbb{E}_{\hat{P}_m}[Z]$, 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 $s_n^2 \geq 2\rho/n \cdot (z_n - \overline{z}_n)^2$, 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

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

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 $\rho$, stepsizes $\rho_t > 0$
$x^{(0)} \leftarrow \mathbf{0}$
for $t = 1, \ldots, T$ do
    $p^{(t)} \leftarrow \operatorname{argmin}_{p \in \mathcal{P}_{\rho,n}} \sum_{i=1}^{n} p_i F_i(x^{(t-1)})$
    Draw $i_1, \ldots, i_c$ from $\{1, \ldots, n\}$
    $\tilde{\nabla}^{(t)} \leftarrow \frac{1}{c} \sum_{\ell=1}^{c} p_{i_\ell}^{(t)} \nabla F_{i_\ell}(x^{(t-1)})$
    $d^{(t)} \leftarrow (1 - \rho_t) d^{(t-1)} + \rho_t \tilde{\nabla}^{(t)}$
    $v^{(t)} \leftarrow \operatorname{argmax}_{v \in \mathcal{X}} \langle d^{(t)}, v \rangle$
    $x^{(t)} \leftarrow x^{(t-1)} + v^{(t)}/T$
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 every 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 suboptimal 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 $h(z) = \min\limits_{p\in\mathcal{P}_{\rho,n}} \langle z, p\rangle$ 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^i_j 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.

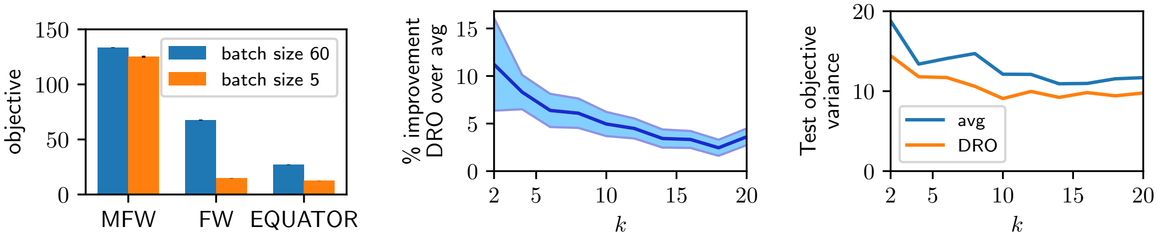
Optimization. We first compare MFW to previously proposed robust optimization algorithms, applied to the DRO problem with $k = 3$. Figure 1a compares **1.** MFW, **2.** Frank-Wolfe (FW) with no momentum and **3.** EQUATOR (Wilder, 2018). Naive FW handles noisy gradients poorly (especially with small 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

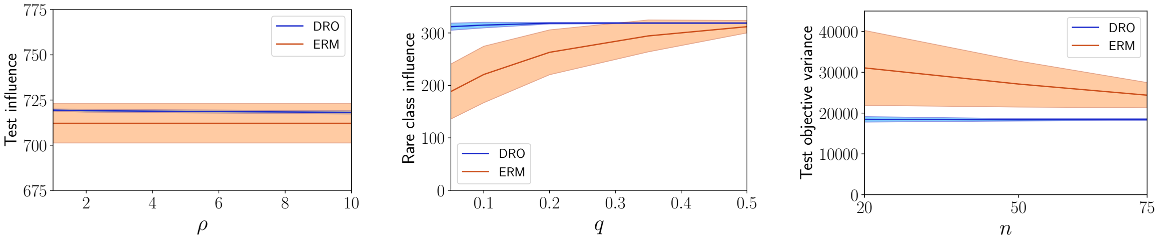
As described in Section 2, we study an influence maximization problem where we observe samples of live-edge graphs $\mathcal{E}_1, \dots, \mathcal{E}_n \sim P$. Our setting is challenging for learning: the number of samples is small and P has high variance. Specifically, P is a mixture of two different 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 nonconvex 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.

A Tail Bound

We use the following one-sided Bernstein's inequality:

Lemma A.1 (Wainwright (2017), Chapter 2)**.** *Let $X_1...X_n$ be iid realizations of a random variable X which satisfies $X \leq B$ almost surely. We have*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \geq \epsilon\right] \leq \exp\left(-\frac{n\epsilon^2}{\mathrm{Var}\,(X) + \frac{B\epsilon}{3}}\right)$$

We apply Lemma A.1 with $X_i = f_i(S)$. If we set the probability on the right hand side to be at most δ, then a simple calculation shows that it suffices to have $n = \frac{\mathrm{Var}(X)}{\epsilon^2}\log\frac{1}{\delta} + \frac{B\epsilon}{3}\log\frac{1}{\delta}$. Hence, for a given value of n, we can guarantee error of at most

$$\epsilon = \sqrt{2\log\left(\frac{1}{\delta}\right)\frac{\mathrm{Var}(X)}{n}} + \frac{2}{3}\log\left(\frac{1}{\delta}\right)\frac{B}{n}.$$

Therefore, we can take $C_1 = \sqrt{2\log\frac{1}{\delta}}$ and $C_2 = \frac{2B}{3}\log\frac{1}{\delta}$. B is often bounded in terms of the problem size for natural submodular maximization problems. For instance, for influence maximization problems we always have $B \leq |V|$ (though tighter bounds may be available for specific graphs and distributions).

B Equivalence of Variance Regularization and Distributionally Robust Optimization

Lemma B.1. *Suppose that $f(\{i\}) \leq b$ for all f in the support of P and all $i \in V$. Then, for each such f, its multilinear extension F is b-Lipschitz in the ℓ_1 norm.*

Proof. Consider any two points $x, x' \in [0,1]^{|V|}$ and any function $f \in \mathrm{support}(P)$. Without loss of generality, let $f(x') \geq f(x)$. Let $[x]^+ = \max(x, 0)$ elementwise, $\vee$ denote elementwise minimum, and 1_i be the vector with a 1 in coordinate i and zeros elsewhere. We bound $F(x')$ as

$$\begin{aligned}
F(x') &\leq F(x' \vee x)\\
&= F(x + [x' - x]^+)\\
&\leq F(x) + F([x' - x]^+)\\
&\leq F(x) + \sum_{i=1}^{|V|} F([x' - x]_i^+ 1_i)\\
&= F(x) + \sum_{i=1}^{|V|} f(\{i\})[x' - x]_i^+\\
&\leq F(x) + b\sum_{i=1}^{|V|}[x' - x]_i^+\\
&\leq F(x) + b\|x' - x\|_1.
\end{aligned}$$

Here, the first inequality follows from monotonicity, while the third and fourth lines use the fact that submodular functions are subadditive, i.e., $F(x + y) \leq F(x) + F(y)$. Now rearranging gives $|F(x') - F(x)| \leq b\|x - x'\|_1$ as desired. □

We will use the following concentration result for the sample variance of a random variable:

Lemma B.2 (Namkoong & Duchi (2017), Section A.1)**.** *Let Z be a random variable bounded in $[0, B]$ and $z_1 ... z_n$ be iid realizations of Z with $n \geq 64$. Let σ denote $\operatorname{Var}(Z)$ and s_n denote the sample variance. It holds that $s_n^2 \geq \frac{1}{4}\sigma^2$ with probability at least $1 - \exp\left(-\frac{n\sigma^2}{36B^2}\right)$.*

This allows us to get a uniform result for the variance expansion of the distributionally robust objective:

Corollary B.1. *Let $\mathcal{X}$ be the polytope $\{x \in [0,1]^{|V|} : \sum_{i=1}^{|V|} x_i = k\}$ corresponding to the k-uniform matroid. With probability at least $1 - \delta$, for all $x \in \mathcal{X}$ such that $\operatorname{Var}_{\mathcal{D}}(F(x)) \geq \frac{\max\{\sqrt{\frac{32}{7}\rho B^2}, \sqrt{36B^2\left(\log\left(\frac{1}{\delta}\right) + |V| \log(1+24k)\right)}\}}{\sqrt{n}}$, the variance expansion holds with equality.*

Proof. Let $\mathcal{X}_{\geq\tau} = \{x : \operatorname{Var}_{\mathcal{P}}(F(x)) \geq \tau\}$ be the set of points x with variance at least τ. Let $\mathcal{Y}$ be a minimal ℓ_1-cover of $\mathcal{X}_{\geq\tau}$ with fineness $\frac{\epsilon}{b}$, for a parameter ϵ to be fixed later. Since the ℓ_1-diameter of $\mathcal{X}$ is $2k$ (by definition), we know that $|\mathcal{Y}| \leq \left(1 + \frac{2kb}{\epsilon}\right)^{|V|}$. Let $s_n(x)$ be the sample variance of $F_1(x), \ldots, F_n(x)$ and $\sigma(x) = \operatorname{Var}_{\mathcal{P}}(F(x))$. Via Lemma B.2 and union bound, we have

$$\Pr\left[s_n^2(x) \geq \frac{1}{4}\sigma^2(x) \ \forall x \in \mathcal{Y}\right] \geq 1 - |\mathcal{Y}| \exp\left(-\frac{n\tau^2}{36B^2}\right).$$

Conditioning on this event, we now extend the sample variance lower bound to the entirety of $\mathcal{X}_{\geq\tau}$. Consider any $x \in \mathcal{X}_{\geq\tau}$ and let $x' \in \arg\min_{x' \in \mathcal{Y}} \|x - x'\|_1$. By definition of $\mathcal{Y}$, $\|x - x\|_1 \leq \frac{\epsilon}{b}$, and so by Lemma B.1, which guarantees Lipschitzness of each F_i, we have $|F_i(x) - F_i(x')| \leq \epsilon$ for all $i = 1, \ldots, n$. Accordingly, it can be shown that $|s_n(x) - s_n(x')| \leq \epsilon$ and $|\sigma(x) - \sigma(x')| \leq \epsilon$. Therefore, we have $s_n(x) \geq s_n(x') - \epsilon \geq \frac{1}{2}\sigma(x') - \epsilon \geq \frac{1}{2}\sigma(x) - \frac{3}{2}\epsilon$. Now by setting $\epsilon = \frac{\tau}{24}$ we have that (conditioned on the above event), $s_n(x) \geq \frac{7}{16}\tau$. Now suppose that we would like the exact variance expansion to hold on all elements of $\mathcal{X}_{\geq\tau}$ with probability at least $1 - \delta$. To have sufficiently high population variance, we must take $\tau \geq \sqrt{\frac{16}{7} \cdot \frac{2\rho B^2}{n}}$. In order for the concentration bound to hold, a simple calculation shows that $\tau \geq \sqrt{\frac{36B^2\left(\log\left(\frac{1}{\delta}\right) + |V| \log(1+24k)\right)}{n}}$ suffices. Taking the max, we need $\tau \geq \frac{\max\{\sqrt{\frac{32}{7}\rho B^2}, \sqrt{36B^2\left(\log\left(\frac{1}{\delta}\right) + |V| \log(1+24k)\right)}\}}{\sqrt{n}}$.

□

C Exact Linear Oracle

In this section we show how to construct a $O(n \log n)$ time *exact* oracle for linear optimization in the χ^2 ball:

$$\begin{array}{ll} \min_p & \langle z, p \rangle \\ \text{s.t.} & \frac{1}{2}\|np - \mathbf{1}\|_2^2 \leq \rho \\ & \mathbf{1}^T p = 1 \\ & p_i \geq 0, \ i = 1, \ldots, n. \end{array} \tag{8}$$

Without loss of generality, assume $z_1 \leq z_2 \leq \cdots \leq z_n$. This can be done by sorting in $O(n \log n)$ time.

First, we wish to discard the case where the χ^2 constraint is not tight. Let k be the largest integer so that $z_1 = z_k$, i.e. $z_1 = \cdots = z_k < z_{k+1}$. If it is feasible, it is optimal to place all the mass of p on the first k coordinates. In particular, the assignment $p_i = 1/k$ for $i = 1, \ldots, k$ accomplishes this while minimizing the χ^2 cost. The cost can be computed as

$$\frac{1}{2}\sum_{i=1}^{k}\left(\frac{n}{k} - 1\right)^2 + \frac{1}{2}\sum_{i=k+1}^{n}(0-1)^2 = \frac{1}{2}\left[k \cdot \left(\frac{n-k}{k}\right)^2 + (n-k)\right] \tag{9}$$

$$= \frac{1}{2} \cdot (n-k) \cdot \left[\frac{n-k}{k} + 1\right] \tag{10}$$

$$= n(n-k)/(2k). \tag{11}$$

Hence if $\rho \geq n(n-k)/(2k)$ we can terminate immediately. Otherwise, we know the χ^2 constraint must be tight.

Before proceeding, we define several auxiliary variables which can all be computed from the problem data in $O(n)$ time:

$$\overline{z}_j = \sum_{i=1}^{j} z_i,\ j = 1, \ldots, n \tag{12}$$

$$b_j = \sum_{i=1}^{j} z_i^2,\ j = 1, \ldots, n \tag{13}$$

$$s_j^2 = \frac{b_j}{j} - (\overline{z}_j)^2,\ j = 1, \ldots, n. \tag{14}$$

Note that $\overline{z}_j$ and s_j^2 are the mean and variance of $\{z_1, \ldots, z_j\}$.

We begin by writing down the Lagrangian of problem (8):

$$\mathcal{L}(p, \lambda, \theta, \eta) = \langle z, p \rangle + \lambda \left(\frac{1}{2} \| np - \mathbf{1} \|_2^2 - \rho \right) + \theta \left(\sum_{i=1}^{n} p_i - 1 \right) - \langle \eta, p \rangle, \tag{15}$$

with dual variables $\lambda \in \mathbb{R}_+$, $\theta \in \mathbb{R}$, and $\eta \in \mathbb{R}_+^n$. By KKT conditions we have

$$0 = \nabla_p \mathcal{L}(p, \lambda, \theta, \eta) = z + \lambda n(np - \mathbf{1}) + \theta \mathbf{1} - \eta. \tag{16}$$

Equivalently,

$$\lambda n^2 p_i = \lambda n - z_i - \theta + \eta_i. \tag{17}$$

By complementary slackness, either $\eta_i > 0$ in which case $p_i = 0$, or we have $\eta_i = 0$ and

$$\lambda n^2 p_i = \lambda n - z_i - \theta. \tag{18}$$

Since $z_1 \leq \cdots \leq z_n$, it follows that p_i decreases as i increases until eventually $p_i = 0$. Hence there exists m so that for $i = 1, \ldots, m$ we have $p_i > 0$ and thereafter $p_i = 0$. Solving for p_i, we have that: for $i = 1, \ldots, m$,

$$p_i = \left(1 - \frac{(z_i + \theta)}{\lambda n} \right) \cdot \frac{1}{n} \text{ for } i = 1, \ldots, m, \tag{19}$$

$$\text{and } p_i = 0 \text{ otherwise.} \tag{20}$$

Note we can divide by λ as we have already determined the corresponding constraint is tight (hence $\lambda > 0$).

We will search for the best choice of m, and then determine p based on m. For fixed λ, m we now solve for the appropriate value of θ. Namely, we must have $\mathbf{1}^T p = 1$:

$$1 = \sum_{i=1}^{n} p_i = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \left(1 - \frac{(z_i + \theta)}{\lambda n} \right) \cdot \frac{1}{n}. \tag{21}$$

Simplifying,

$$n = \sum_{i=1}^{m} \left(1 - \frac{(z_i + \theta)}{\lambda n} \right) = m - \frac{1}{\lambda n} \sum_{i=1}^{m} (z_i + \theta) \tag{22}$$

$$= m - \frac{m \overline{z}_m}{\lambda n} - \frac{\theta m}{\lambda n}. \tag{23}$$

Multiplying through by λn and solving for θ, we have

$$\lambda n^2 = \lambda m n - m \overline{z}_m - \theta m \implies \theta = \left(1 - \frac{n}{m} \right) \lambda n - \overline{z}_m. \tag{24}$$

Now that we have solved for θ as a function of λ and m, the variable p is purely a function of m and λ. For fixed λ and m, it is not hard to compute the objective value attained by the value of p induced by equation (19):

$$\begin{aligned}
\langle z, p \rangle &= \frac{1}{n}\sum_{i=1}^{m}\left(1 - \frac{(z_i+\theta)}{\lambda n}\right) z_i && (25)\\
&= \frac{1}{n}\sum_{i=1}^{m} z_i - \frac{1}{n}\sum_{i=1}^{m}\frac{(z_i+\theta)z_i}{\lambda n} && (26)\\
&= \frac{m}{n}\overline{z}_m - \frac{1}{\lambda n^2}\sum_{i=1}^{m}(z_i^2 + \theta z_i) && (27)\\
&= \frac{m}{n}\overline{z}_m - \frac{1}{\lambda n^2}(b_m + \theta m \overline{z}_m) && (28)\\
&= \frac{m}{n}\overline{z}_m - \frac{1}{\lambda n^2}\left(b_m + \left(\left(1 - \frac{n}{m}\right)\lambda n - \overline{z}_m\right) m\overline{z}_m\right) && (29)\\
&= \frac{m}{n}\overline{z}_m - \frac{b_m}{\lambda n^2} - \frac{(1-n/m)\lambda n m \overline{z}_m}{\lambda n^2} + \frac{m(z_m)^2}{\lambda n^2} && (30)\\
&= \frac{m}{n}\overline{z}_m - \frac{b_m}{\lambda n^2} + \frac{(n-m)\overline{z}_m}{n} + \frac{m(z_m)^2}{\lambda n^2} && (31)\\
&= \overline{z}_m - \frac{b_m}{\lambda n^2} + \frac{m(z_m)^2}{\lambda n^2} && (32)\\
&= \overline{z}_m - \frac{1}{\lambda n^2}\cdot(b_m - m(\overline{z}_m)^2) && (33)\\
&= \overline{z}_m - \frac{m s_m^2}{\lambda n^2}. && (34)
\end{aligned}$$

Since $ms_m^2 \geq 0$, for fixed m we seek the minimum value of λ such that the induced p is still feasible. Since the $\mathbf{1}^T p = 1$ constraint is guaranteed by the choice of θ, we need only check the χ^2 and nonnegativity constraints. In section C.1 we derive that the optimal feasible λ is given by

$$\lambda = \frac{1}{n^2}\cdot\max\left\{\sqrt{\frac{m^2 s_m^2}{\alpha(m,n,\rho)}},\ m(z_m - \overline{z}_m)\right\}. \tag{35}$$

Hence, in constant time for each candidate m with $\alpha(m, n, \rho)$, we select λ per equation (35) and evaluate the objective. Finally, we return p corresponding to the optimal choice m. This algorithm is given more formally in Algorithm 2.

C.1 Constraints on λ for fixed m

First we check the χ^2 constraint; since $\lambda > 0$, we have:

$$\begin{aligned}
\rho &\geq \frac{1}{2}\|np - \mathbf{1}\|_2^2 && (36)\\
&= \frac{1}{2}\sum_{i=1}^{n}(np_i - 1)^2 && (37)\\
&= \frac{1}{2}\sum_{i=1}^{m}\left(\left(1 - \frac{(z_i+\theta)}{\lambda n}\right) - 1\right)^2 + \frac{1}{2}\sum_{i=m+1}^{n}(-1)^2 && (38)\\
&= \frac{1}{2}\cdot\frac{1}{\lambda^2 n^2}\sum_{i=1}^{m}(z_i+\theta)^2 + \frac{1}{2}(n-m). && (39)
\end{aligned}$$

Algorithm 2 Linear optimization in χ^2 ball

```
  Input: pre-sorted vector $z$ with $z_1 \le \cdots \le z_n$
  Output: optimal vector $p$
  Compute maximum $k$ s.t. $z_1 = z_k$
  if $n(n-k)/(2k) \le \rho$ then
    return $p$ with $p_i = 1/k \cdot \mathbf{1}\{i \le k\}$
  end if
  $\{$now we must search for optimal $m\}$
  $\overline{z}_j \leftarrow \frac{1}{j}\sum_{i=1}^{j} z_i,\ j = 1, \ldots, n$
  $b_j \leftarrow \sum_{i=1}^{j} z_i^2,\ j = 1, \ldots, n$
  $s_j^2 \leftarrow b_j/j - (\overline{z}_j)^2,\ j = 1, \ldots, n$
  $m_{\min} \leftarrow \min\{m \in \{1, \ldots, n\} : \alpha(m, n, \rho) > 0\}$
  $\lambda_m = \frac{1}{n^2} \cdot \max\left\{\sqrt{\frac{m^2 s_m^2}{\alpha(m,n,\rho)}},\ (z_m - \overline{z}_m)m\right\},\ m = m_{\min}, \ldots, n$
  $v_m \leftarrow \overline{z}_m - m s_m^2/(\lambda_m n^2),\ m = m_{\min}, \ldots, n$
  $m_{\text{opt}} \leftarrow \operatorname{argmin}_m\{v_m : m = m_{\min}, \ldots, n\}$
  $\theta \leftarrow \left(1 - \frac{n}{m_{\text{opt}}}\right)\lambda_{m_{\text{opt}}} n - \overline{z}_{m_{\text{opt}}}$
  return $p = \frac{1}{n}\max\left(0, 1 - \frac{z_{m_{\text{opt}}} + \theta}{\lambda_{m_{\text{opt}}} n}\right)$
```

We expand the sum of squares:

$$\begin{aligned}
\sum_{i=1}^{m}(z_i + \theta)^2 &= \sum_{i=1}^{m}(z_i^2 + 2z_i\theta + \theta^2) && (40)\\
&= \sum_{i=1}^{m} z_i^2 + 2\theta\sum_{i=1}^{m} z_i + \sum_{i=1}^{m}\theta^2 && (41)\\
&= b_m + 2\theta m\overline{z}_m + \theta^2 m. && (42)
\end{aligned}$$

Plugging in our expression for θ, this equals:

$$\begin{aligned}
b_m + 2\theta m\overline{z}_m + \theta^2 m &= b_m + 2m\overline{z}_m\theta + \left[\left(1 - \frac{n}{m}\right)\lambda n - \overline{z}_m\right]^2 \cdot m && (43)\\
&= b_m + 2m\overline{z}_m\theta + \left[\left(1 - \frac{n}{m}\right)^2 \lambda^2 n^2 - 2\left(1 - \frac{n}{m}\right)\lambda n \cdot \overline{z}_m + (\overline{z}_m)^2\right] \cdot m && (44)\\
&= b_m + 2m\overline{z}_m\theta + \left(1 - \frac{n}{m}\right)^2 \lambda^2 n^2 m - 2\left(1 - \frac{n}{m}\right)\lambda n m\overline{z}_m + m(\overline{z}_m)^2 && (45)\\
&= b_m + 2m\overline{z}_m\left[\left(1 - \frac{n}{m}\right)\lambda n - \overline{z}_m\right] + \left(1 - \frac{n}{m}\right)^2 \lambda^2 n^2 m - 2\left(1 - \frac{n}{m}\right)\lambda n m\overline{z}_m + m(\overline{z}_m)^2 && (46)\\
&= b_m + 2\left(1 - \frac{n}{m}\right)\lambda n m\overline{z}_m - 2m(\overline{z}_m)^2 + \left(1 - \frac{n}{m}\right)^2 \lambda^2 n^2 m - 2\left(1 - \frac{n}{m}\right)\lambda n m\overline{z}_m + m(\overline{z}_m)^2 && (47)\\
&= b_m - 2m(\overline{z}_m)^2 + \left(1 - \frac{n}{m}\right)^2 \lambda^2 n^2 m + m(\overline{z}_m)^2 && (48)\\
&= b_m - m(\overline{z}_m)^2 + \left(1 - \frac{n}{m}\right)^2 \lambda^2 n^2 m && (49)\\
&= m s_m^2 + \left(1 - \frac{n}{m}\right)^2 \lambda^2 n^2 m. && (50)
\end{aligned}$$

Finally, plugging this back into equation (39) yields:

$$\rho \geq \frac{1}{2} \cdot \frac{1}{\lambda^2 n^2} \cdot \left[m s_m^2 + \left(1 - \frac{n}{m}\right)^2 \lambda^2 n^2 m \right] + \frac{1}{2} \cdot (n - m) \tag{51}$$

$$\Leftrightarrow 2\rho \geq \frac{m s_m^2}{\lambda^2 n^2} + \left(1 - \frac{n}{m}\right)^2 m + (n - m) \tag{52}$$

$$\Leftrightarrow 2\rho \geq \frac{m s_m^2}{\lambda^2 n^2} + \left(1 - \frac{2n}{m} + \frac{n^2}{m^2}\right) m + (n - m) \tag{53}$$

$$\Leftrightarrow 2\rho \geq \frac{m s_m^2}{\lambda^2 n^2} + m - 2n + \frac{n^2}{m} + (n - m) \tag{54}$$

$$\Leftrightarrow 2\rho \geq \frac{m s_m^2}{\lambda^2 n^2} - n + \frac{n^2}{m} \tag{55}$$

$$\Leftrightarrow \frac{2\rho m}{n^2} \geq \frac{m^2 s_m^2}{\lambda^2 n^4} - \frac{m}{n} + 1 \tag{56}$$

$$\Leftrightarrow \frac{m^2 s_m^2}{\lambda^2 n^4} \leq \alpha(m, n, \rho), \tag{57}$$

where $\alpha(m, n, \rho)$ is defined as in the main text. If $\alpha(m, n, \rho) \leq 0$, there is no feasible choice of λ for this m. Otherwise, we can divide and solve for λ:

$$\lambda \geq \sqrt{\frac{m^2 s_m^2}{n^4 \alpha(m, n, \rho)}} = \frac{1}{n} \sqrt{\frac{m s_m^2}{2\rho + n - n^2/m}}, \tag{58}$$

or equivalently

$$\lambda n^2 \geq \sqrt{\frac{m^2 s_m^2}{\alpha(m, n, \rho)}}. \tag{59}$$

Now we check the other remaining constraint on λ, that the constraint $p_i \geq 0$ for $i = 1, \ldots, m$ must hold. In particular, we must have $p_m \geq 0$:

$$0 \leq p_m = \frac{1}{n} \cdot \left(1 - \frac{z_m + \theta}{\lambda n}\right) \tag{60}$$

$$\Leftrightarrow z_m + \theta \leq \lambda n \tag{61}$$

$$\Leftrightarrow z_m + \left(1 - \frac{n}{m}\right) \lambda n - \overline{z}_m \leq \lambda n \tag{62}$$

$$\Leftrightarrow z_m - \overline{z}_m \leq \frac{\lambda n^2}{m} \tag{63}$$

$$\Leftrightarrow m(z_m - \overline{z}_m) \leq \lambda n^2. \tag{64}$$

Hence λ must satisfy

$$\lambda n^2 \geq \max \left\{ \sqrt{\frac{m^2 s_m^2}{\alpha(m, n, \rho)}},\ m(z_m - \overline{z}_m) \right\}. \tag{65}$$

Since we seek minimal λ, we select λ which makes this constraint tight.

C.2 Unique solutions

Here we provide results for understanding when there is a unique solution to Problem (8). Recall that our solution to Problem (8) first checks whether the optimal solutions have tight χ^2 constraint. By choosing ρ small enough, this can be guaranteed uniformly:

Lemma C.1. *Suppose $\{z_i\}$ attain at least ℓ distinct values. If $\rho \leq (\ell - 1)/2$ then all optimal solutions to Problem (8) have tight χ^2 constraint.*

Proof. Assume $z_1 \leq \cdots \leq z_n$. If $\{z_i\}$ attain at least ℓ distinct values, then the maximum number k so that $z_1 = \cdots = z_k$ can be bounded by $n - \ell + 1$. Recall from earlier in section C that the constraint is tight if $\rho \leq n(n-k)/(2k)$, and note that this bound is monotone decreasing in k. Hence, we can guarantee the constraint is tight as long as

$$\rho \leq \frac{n(n-(n-\ell+1))}{2(n-\ell+1)} = \frac{n(\ell-1)}{2(n-\ell+1)}. \tag{66}$$

Since $n-\ell+1 \leq n$, the previous inequality is implied by

$$\rho \leq \frac{(n-\ell+1)(\ell-1)}{2(n-\ell+1)} = \frac{\ell-1}{2}. \tag{67}$$

□

Now, assuming the χ^2 constraint is tight, we can characterize the set of optimal solutions:

Lemma C.2. *Suppose the optimal solutions for Problem (8) all have tight χ^2 constraint. Then there is a unique optimal solution p^* with minimum cardinality among all optimal solutions.*

Proof. This is a consequence of our characterization of the optimal dual variable λ as a function of the sparsity m. For each choice of m, we solved earlier for the unique dual variable λ_m which determines a unique solution p. Hence, even if there are multiple values of m that are feasible and that yield optimal objective value, there is still a unique minimal m_{opt}, which in turn yields a unique optimal solution. □

C.3 Lipschitz gradient

Lemma C.3. *Define $h(z) = \min_{p \in \mathcal{P}_{\rho,n}} \langle z, p \rangle$. Then on the subset of z's satisfying the high sample variance condition $s_n^2 \geq (2\rho B^2)/n^2$, $h(z)$ has Lipschitz gradient with constant $L \leq \frac{2\sqrt{2\rho}}{n^{3/2}} + \frac{2}{Bn^{1/2}}$.*

Proof. In this regime, there is a unique worst-case $p \in \mathcal{P}_{\rho,n}$, and it is the gradient of $h(z)$. In the high sample variance regime, we have $m = n$, i.e. each $p_i > 0$ and:

$$p_i = \left(1 - \frac{z_i + \theta}{\lambda n}\right) \cdot \frac{1}{n} \text{ for all } i = 1, \dots, n. \tag{68}$$

In particular, $\theta = (1 - n/n)\lambda n - \overline{z}_n = -\overline{z}_n$, and $\lambda = \frac{1}{n^2}\sqrt{n^2 s_n^2/(2\rho/n)}$. Simplifying, we have

$$\begin{aligned} p_i &= \left(1 - \frac{z_i - \overline{z}_n}{\lambda n}\right) \cdot \frac{1}{n} && (69)\\ &= \left(1 - \frac{z_i - \overline{z}_n}{\frac{1}{n}\sqrt{n^2 s_n^2/(2\rho/n)}}\right) \cdot \frac{1}{n} && (70)\\ &= \left(1 - \frac{z_i - \overline{z}_n}{\sqrt{n s_n^2/(2\rho)}}\right) \cdot \frac{1}{n}. && (71) \end{aligned}$$

We will bound the Lipschitz constant of p as a function of z by computing the Hessian which has entries $H_{ij} = \frac{\partial p_i}{\partial z_j}$ and bounding its largest eigenvalue. For the element H_{ij} we have two cases. If $i = j$, then

$$\begin{aligned} H_{ii} &= -\frac{\sqrt{2\rho}}{n^{3/2}} \cdot \frac{\partial}{\partial z_i}\left(\frac{z_i - \overline{z}_n}{\sqrt{s_n^2}}\right) && (72)\\ &= -\frac{\sqrt{2\rho}}{n^{3/2}} \cdot \left(\frac{\sqrt{s_n^2}(1 - \frac{1}{n}) - (z_i - \overline{z}_n) \cdot \frac{2}{n} \cdot (z_i - \overline{z}_n)}{s_n^2}\right). && (73) \end{aligned}$$

If $i \neq j$, then

$$\begin{aligned} H_{ij} &= -\frac{\sqrt{2\rho}}{n^{3/2}} \cdot \frac{\partial}{\partial z_j}\left(\frac{z_i - \overline{z}_n}{\sqrt{s_n^2}}\right) && (74)\\ &= -\frac{\sqrt{2\rho}}{n^{3/2}} \cdot \left(\frac{-\frac{1}{n} \cdot \sqrt{s_n^2} - (z_i - \overline{z}_n) \cdot \frac{2}{n} \cdot (z_j - \overline{z}_n)}{s_n^2}\right). && (75) \end{aligned}$$

Define $\tilde{H}$ so that $\frac{\sqrt{2\rho}}{n^{3/2}s_n^2}\tilde{H} = H$, i.e.

$$\tilde{H}_{ij} = \begin{cases} \sqrt{s_n^2}(\frac{1}{n} - 1) + (z_i - \overline{z}_n) \cdot \frac{2}{n} \cdot (z_i - \overline{z}_n) & i = j \\ \frac{1}{n} \cdot \sqrt{s_n^2} + (z_i - \overline{z}_n) \cdot \frac{2}{n} \cdot (z_j - \overline{z}_n) & i \neq j. \end{cases} \tag{76}$$

It is easy to see that $\tilde{H}$ is given by

$$\tilde{H} = -\text{diag}(\sqrt{s_n^2}\mathbf{1}) + \frac{\sqrt{s_n^2}}{n}\mathbf{1}\mathbf{1}^T + \frac{2}{n}(z - \overline{z}_n\mathbf{1})(z - \overline{z}_n\mathbf{1})^T. \tag{77}$$

By the triangle inequality, the operator norm of $\tilde{H}$ can thus be bounded by

$$\begin{aligned} \|\tilde{H}\| &\leq \|\text{diag}(\sqrt{s_n^2}\mathbf{1})\| + \frac{\sqrt{s_n^2}}{n}\|\mathbf{1}\mathbf{1}^T\| + \frac{2}{n}\|(z - \overline{z}_n\mathbf{1})(z - \overline{z}_n\mathbf{1})^T\| & (78) \\ &= \sqrt{s_n^2} + \frac{\sqrt{s_n^2}}{n}\|\mathbf{1}\|_2^2 + \frac{2}{n}\|z - \overline{z}_n\mathbf{1}\|_2^2 & (79) \\ &= 2\sqrt{s_n^2} + \frac{2}{n}\sum_{i=1}^n (z_i - \overline{z}_n)^2 & (80) \\ &= 2\sqrt{s_n^2} + 2s_n^2. & (81) \end{aligned}$$

It follows that the Lipschitz constant of the gradient of $h(z)$ can be bounded by

$$\begin{aligned} \|H\| &= \frac{\sqrt{2\rho}}{n^{3/2}s_n^2}\|\tilde{H}\| & (82) \\ &\leq \frac{\sqrt{2\rho}}{n^{3/2}s_n^2}\left(2\sqrt{s_n^2} + 2s_n^2\right) & (83) \\ &= \frac{2\sqrt{2\rho}}{n^{3/2}} \cdot \left(1 + \frac{1}{\sqrt{s_n^2}}\right). & (84) \end{aligned}$$

Since we are in the high variance regime $s_n^2 \geq (2\rho B^2)/n$, it follows that $1/\sqrt{s_n^2} \leq \sqrt{n}/(B\sqrt{2\rho})$ and therefore

$$\begin{aligned} \|H\| &\leq \frac{2\sqrt{2\rho}}{n^{3/2}} \cdot \left(1 + \frac{\sqrt{n}}{B\sqrt{2\rho}}\right) & (85) \\ &= \frac{2\sqrt{2\rho}}{n^{3/2}} + \frac{2}{Bn}. & (86) \end{aligned}$$

□

D Projection onto the χ^2 ball

Let $w \in \mathbb{R}^n$ be pre-sorted (taking time $O(n \log n)$), so that $w_1 \geq \cdots \geq w_n$. We wish to solve the problem

$$\begin{array}{ll} \min_p & \frac{1}{2}\|p - w\|_2^2 \\ \text{s.t.} & \frac{1}{2}\|np - \mathbf{1}\|_2^2 \leq \rho \\ & \mathbf{1}^T p = 1 \\ & p_i \geq 0, \ i = 1, \ldots, n. \end{array} \tag{87}$$

As in section C, we start by precomputing the auxiliary variables:

$$\overline{w}_j = \sum_{i=1}^j w_i, \ j = 1, \ldots, n \tag{88}$$

$$b_j = \sum_{i=1}^j w_i^2, \ j = 1, \ldots, n \tag{89}$$

$$s_j^2 = \frac{b_j}{j} - (\overline{w}_j)^2, \ j = 1, \ldots, n. \tag{90}$$

We remark that these can be updated efficiently when sparse updates are made to w; coupled with a binary search over optimal m, this can yield $O(\log n)$ update time as in (Duchi et al., 2008; Namkoong & Duchi, 2016).

We form the Lagrangian:

$$\mathcal{L}(p, \lambda, \theta, \eta) = \frac{1}{2}\|p - w\|_2^2 + \lambda\left(\frac{1}{2}\|np - \mathbf{1}\|_2^2 - \rho\right) + \theta\left(\sum_{i=1}^{n} p_i - 1\right) - \langle \eta, p\rangle \tag{91}$$

with dual variables $\lambda \in \mathbb{R}_+, \theta \in \mathbb{R}$, and $\eta \in \mathbb{R}_+^n$. We will also use the reparameterization $\beta = 1/(1 + \lambda n^2)$ throughout. By KKT conditions we have

$$\begin{aligned} 0 = \nabla_p \mathcal{L}(p, \lambda, \theta, \eta) &= p - w + \lambda n\,(np - \mathbf{1}) + \theta\mathbf{1} - \eta && (92)\\ &= (1 + \lambda n^2)p - w - \lambda n + \theta\mathbf{1} - \eta. && (93)\end{aligned}$$

For any given i, if $\eta_i > 0$ we have $p_i = 0$ by complementary slackness. Otherwise, if $\eta_i = 0$ we have

$$\begin{aligned} 0 &= (1 + \lambda n^2)p_i - w_i - \lambda n + \theta && (94)\\ \Leftrightarrow (1 + \lambda n^2)p_i &= w_i + \lambda n - \theta. && (95)\end{aligned}$$

The variable p is implicitly given here by θ, λ and m. Next we seek to solve for θ as a function of λ and m.

Note that since w_i decreases as i increases, therefore p_i also decreases. It follows that for some $m \in \{1, \ldots, n\}$, we have $p_i > 0$ for $i \leq m$ and $p_i = 0$ otherwise. Since p_i must sum to one, we have

$$\begin{aligned} 1 &= \sum_{i=1}^{n} p_i = \sum_{i=1}^{m} p_i && (96)\\ \Leftrightarrow (1 + \lambda n^2) &= \sum_{i=1}^{m}(w_i + \lambda n - \theta) && (97)\\ &= m\overline{w}_m + \lambda mn - m\theta && (98)\end{aligned}$$

from which it follows that $\theta = \overline{w}_m + \lambda n - (1 + \lambda n^2)/m$. Plugging this into the expression for p_i and rearranging yields

$$p_i = (w_i - \overline{w}_m)\beta + 1/m. \tag{99}$$

It will become apparent later that the objective improves as β increases, and so for fixed m we seek the largest β which yields a feasible p. First, we check the χ^2 constraint:

$$\begin{aligned} \rho &\geq \frac{1}{2}\sum_{i=1}^{n}(np_i - 1)^2 && (100)\\ &= \frac{1}{2}\sum_{i=1}^{m}(np_i - 1)^2 + \frac{1}{2}\sum_{i=m+1}^{n} 1^2 && (101)\\ &= \frac{1}{2}\sum_{i=1}^{m}(n\beta(w_i - \overline{w}_m) + n/m - 1)^2 + \frac{1}{2}(n - m). && (102)\end{aligned}$$

Expanding and multiplying by 2, we have

$$2\rho \geq \sum_{i=1}^{m}\left[n^2\beta^2(w_i - \overline{w}_m)^2 + 2n\beta(n/m - 1)(w_i - \overline{w}_m) + (n/m - 1)^2\right] + n - m. \tag{103}$$

The middle term in the sum cancels because $\sum_{i=1}^{m} w_i = m\overline{w}_m$. We are left with

$$\begin{aligned} 2\rho &\geq n^2\beta^2\sum_{i=1}^{m}(w_i - \overline{w}_m)^2 + m(n/m - 1)^2 + n - m && (104)\\ &= n^2\beta^2 m s_m^2 + m(n/m - 1)^2 + n - m. && (105)\end{aligned}$$

Solving for β^2, we are left with

$$\beta^2 \le \frac{2\rho + n - n^2/m}{n^2 m s_m^2} = \frac{2\rho m/n^2 + m/n - 1}{m^2 s_m^2} = \frac{\alpha(m,n,\rho)}{m^2 s_m^2}, \tag{106}$$

where $\alpha(m,n,\rho)$ is defined as in the main text. This gives the maximum value of β for which the χ^2 constraint is met. We also need to check the $p_i \ge 0$ constraint. This is more straightforward: we must have

$$0 \le p_i = (w_i - \overline{w}_m)\beta + 1/m \tag{107}$$

for all $i = 1, \ldots, m$. Since w_i is decreasing, it suffices to check $i = m$. If $w_m - \overline{w}_m \ge 0$ there is no problem, as $\beta > 0$. Otherwise, we divide and are left with the condition

$$\beta \le \frac{1}{m(\overline{w}_m - w_m)}. \tag{108}$$

Our exact algorithm is now straightforward: for each m, compute the largest feasible β (if there is a feasible β), compute the corresponding objective value, and then return p corresponding to the best m.

If $\alpha(m,n,\rho) < 0$ for a given m, we can immediately discard that choice of m as infeasible. Otherwise we compute β and check the objective value v_m for that m:

$$\begin{aligned}
v_m &= \frac{1}{2}\|p - w\|_2^2 && (109)\\
&= \frac{1}{2}\sum_{i=1}^m ((w_i - \overline{w}_m)\beta + 1/m - w_i)^2 + \frac{1}{2}\sum_{i=m+1}^n w_i^2 && (110)\\
&= \frac{1}{2}\sum_{i=1}^m \left[(w_i - \overline{w}_m)^2\beta^2 + 2\beta(w_i - \overline{w}_m)(1/m - w_i) + (1/m - w_i)^2\right] + \frac{1}{2}\sum_{i=m+1}^n w_i^2. && (111)
\end{aligned}$$

As before, the $\sum_{i=1}^m 2\beta(w_i - \overline{w}_m)/m$ term cancels and we are left with

$$\begin{aligned}
v_m &= \frac{1}{2}\sum_{i=1}^m \left[(w_i - \overline{w}_m)^2\beta^2 - 2\beta w_i(w_i - \overline{w}_m) + (1/m - w_i)^2\right] + \frac{1}{2}\sum_{i=m+1}^n w_i^2 && (112)\\
&= \frac{1}{2}\cdot\beta^2 m s_m^2 - \beta\sum_{i=1}^m w_i^2 + \beta\sum_{i=1}^m w_i\overline{w}_m + \frac{1}{2}\sum_{i=1}^m (1/m - w_i)^2 + \frac{1}{2}\sum_{i=m+1}^n w_i^2 && (113)\\
&= \frac{1}{2}\cdot\beta^2 m s_m^2 - \beta b_m + \beta m(\overline{w}_m)^2 + \frac{1}{2}\sum_{i=1}^m (1/m^2 - 2w_i/m + w_i^2) + \frac{1}{2}(b_n - b_m) && (114)\\
&= \frac{1}{2}\cdot\beta^2 m s_m^2 - \beta b_m + \beta m(\overline{w}_m)^2 + \frac{1}{2}\cdot\frac{1}{m} - \overline{w}_m + \frac{1}{2}b_m + \frac{1}{2}(b_n - b_m) && (115)\\
&= \frac{1}{2}\cdot\beta^2 m s_m^2 - \beta b_m + \beta m(\overline{w}_m)^2 + \frac{1}{2m} - \overline{w}_m + \frac{1}{2}b_n && (116)\\
&= \frac{1}{2}\cdot\frac{\alpha(m,n,\rho)}{m} - \beta(b_m - m(\overline{w}_m)^2) + \frac{1}{2m} - \overline{w}_m + \frac{1}{2}b_n && (117)\\
&= \left(\frac{\rho}{n^2} + \frac{1}{2n} - \frac{1}{2m}\right) - \beta m s_m^2 + \frac{1}{2m} - \overline{w}_m + \frac{1}{2}b_n && (118)\\
&= \left(\frac{\rho}{n^2} + \frac{1}{2n}\right) - \beta m s_m^2 - \overline{w}_m + \frac{1}{2}b_n. && (119)
\end{aligned}$$

Discarding the terms which do not depend on m, we seek m which minimizes $\tilde{v}_m := -\beta m s_m^2 - \overline{w}_m$. Finally, we remark that it is now quite apparent that for fixed m we wish to maximize β.

E Convergence analysis for MFW

Here we establish the convergence rate of the MFW algorithm specifically for the DRO problem. For completeness, we reproduce the MFW convergence guarantee here:

Lemma E.1 (adapted from Mokhtari et al. (2018))**.** *Let F be an up-concave function with L-Lipschitz gradient. MFW run for T iterations returns a point $x^{(T)}$ satisfying*

$$\mathbb{E}[F(x^{(T)})] \geq \left(1 - \frac{1}{e}\right) OPT - \frac{2DQ^{1/2}}{T^{1/3}} - \frac{LD^2}{2T}$$

where $D = \max_{x \in \mathcal{X}} \|x\|$, $Q = \max\{9^{2/3}\|\nabla F(x_0) - d_0\|^2, 16\sigma^2 + 3L^2D^2\}$, and σ is the variance of the stochastic gradients.

The main work is to establish Lipschitz continuity of ∇G, the gradient of the DRO objective. In fact, Mokhtari et al. (2018) get a better bound by controlling changes in ∇G specifically along the updates used by MFW. We bound this same quantity as follows:

Lemma E.2. *When the high sample variance condition is satisfied, for any two points $x^{(t)}$ and $x^{(t+1)}$ produced by MFW, ∇G satisfies $\|\nabla G(x^{(t)}) - \nabla G(x^{(t+1)})\| \leq \left(b\sqrt{n|V|}L + b\sqrt{k}\right)\|x^{(t)} - x^{(t+1)}\|$.*

Proof. We write $\vec{F}(x) = (F_1(x), ..., F_n(x))$, and are interested in the composition $G = h(\vec{F}(x))$ (recall that h is defined in Lemma 3.2 as the value of the inner minimization problem for a given set of values). Let $D\vec{F}(x)$ be the matrix derivative of $\vec{F}$. That is, $\left[D\vec{F}(x)\right]_{ij} = \frac{\partial}{\partial x_j}F_i(x)$. The chain rule yields

$$\nabla h(\vec{F}(x)) = \left(\nabla h(\vec{F}(x))\right) D\vec{F}(x).$$

Consider two points $x, y \in \mathcal{X}$. To apply the argument of Mokhtari et al. (2018), we would like a bound on the change in ∇h along the MFW update from x in the direction of y. Let $x' = x + \frac{1}{T}y$ be the updated point. We have

$$\begin{aligned}
\|\nabla h(\vec{F}(x)) - \nabla h(\vec{F}(x'))\| &= \left\|\left(\nabla h(\vec{F}(x))\right) D\vec{F}(x) - \left(\nabla h(\vec{F}(x'))\right) D\vec{F}(x')\right\| \\
&= \Big\|\left(\nabla h(\vec{F}(x))\right) D\vec{F}(x) - \left(\nabla h(\vec{F}(x))\right) D\vec{F}(x') \\
&\qquad + \left(\nabla h(\vec{F}(x))\right) D\vec{F}(x') - \left(\nabla h(\vec{F}(x'))\right) D\vec{F}(x')\Big\| \\
&\leq \left\|\left(\nabla h(\vec{F}(x))\right) D\vec{F}(x) - \left(\nabla h(\vec{F}(x))\right) D\vec{F}(x')\right\| \\
&\qquad + \left\|\left(\nabla h(\vec{F}(x))\right) D\vec{F}(x') - \left(\nabla h(\vec{F}(x'))\right) D\vec{F}(x')\right\| \\
&= \left\|\left(\nabla h(\vec{F}(x))\right) \left(D\vec{F}(x) - D\vec{F}(x')\right)\right\| \\
&\qquad + \left\|\left(\nabla h(\vec{F}(x)) - \nabla h(\vec{F}(x'))\right) D\vec{F}(x')\right\|.
\end{aligned}$$

Starting out with the first term, we note that $\nabla h(\vec{F}(x))$ is a probability vector (the optimal p for the DRO problem). Hence, we have

$$\begin{aligned}
\left\|\left(\nabla h(\vec{F}(x))\right)\left(D\vec{F}(x) - D\vec{F}(x')\right)\right\| &\leq \max_{i=1...n}\left\|D\vec{F}(x)_i - D\vec{F}(x')_i\right\| \\
&= \max_{i=1...n}\|\nabla F_i(x) - \nabla F_i(x')\|
\end{aligned}$$

And from Lemma 3 of Mokhtari et al. (2018), we have that when x' is an updated point of the MFW algorithm starting at x,

$$\|\nabla F_i(x) - \nabla F_i(x')\| \leq b\sqrt{k}\|x - x'\| \quad \forall i = 1...n.$$

We now turn to the second term. Note that the jth component of this vector is just the dot product

$$\left(\nabla h(\vec{F}(x)) - \nabla h(\vec{F}(x'))\right) \cdot D\vec{F}(x)_{\cdot,j}$$

where $D\vec{F}(x)_{\cdot,j}$ collects the partial derivative of each F_i with respect to x_j. Via the Cauchy-Schwartz inequality, we have

$$\left(\nabla h(\vec{F}(x)) - \nabla h(\vec{F}(x'))\right) \cdot D\vec{F}(x)_{\cdot,j} \leq \left\|\left(\nabla h(\vec{F}(x)) - \nabla h(\vec{F}(x'))\right)\right\| \left\|D\vec{F}(x)_{\cdot,j}\right\|$$

Lemma 3.2 shows that $\left\|\left(\nabla h(\vec{F}(x)) - \nabla h(\vec{F}(x'))\right)\right\| \leq L\|x - x'\|$. In order to bound the second norm, we claim that for all i, j, $\nabla_j F_i(x) \leq b$. To show this, note that we can use the definition of the multilinear extension to write

$$\nabla_j F_i(x) = \mathbb{E}_{S\sim x}[f_i(S|\{j\} \in S)] - \mathbb{E}_{S\sim x}[f_i(S|\{j\} \notin S)]$$

where $S \sim x$ denotes that S is drawn from the product distribution with marginals x. Now it is simple to show using submodularity of f_i that

$$\mathbb{E}_{S\sim x}[f_i(S|\{j\} \in S)] - \mathbb{E}_{S\sim x}[f_i(S|\{j\} \notin S)] \leq f_i(\{j\}) - f_i(\emptyset) \leq b.$$

Accordingly, we have that

$$\left\|D\vec{F}(x)_{\cdot,j}\right\| \leq b\|\mathbf{1}\| = b\sqrt{n}.$$

This gives us a component-wise bound on each element of the vector $\left(\nabla h(\vec{F}(x)) - \nabla h(\vec{F}(x'))\right) D\vec{F}(x')$. Putting it all together, we have

$$\begin{aligned} \left\|\left(\nabla h(\vec{F}(x)) - \nabla h(\vec{F}(x'))\right) D\vec{F}(x')\right\| &\leq b\sqrt{n}L\|x - x'\| \cdot \|\mathbf{1}\| \\ &\leq b\sqrt{n|V|} \cdot L \cdot \|x - x'\|, \end{aligned}$$

and summing the two terms yields the final Lipschitz constant $b\sqrt{n|V|}L + b\sqrt{k}$. □

Now the final convergence rate for MFW stated in Theorem 3.1 follows from plugging the above Lipschitz bound into Lemma E.1. We also remark that the above argument trivially goes through for an arbitrary (not necessarily submodular) functions:

Lemma E.3. *Suppose that each function $f : \mathbb{R}^{|V|} \to \mathbb{R}$ in the support of P has bounded norm gradients $\max_{i=1\ldots|V|} |\nabla_i f| \leq b$ which are also L_f-Lipschitz. Then under the high variance condition, the corresponding DRO objective G has L_G-Lipschitz gradient with $L_G \leq L_f + b\sqrt{n|V|}L$, where L is as defined in Lemma 3.2.*

F Rounding to a distribution over subsets

The output of MFW is a fractional vector $x \in \mathcal{X}$. Lemma 3.1 guarantees this x can be converted into a distribution $\mathcal{D}$ over feasible subsets, and moreover, that the attainable solution value from doing so is within a $(1-1/e)$ factor of the optimal value for the DRO problem. This result is essentially standard (see Wilder (2018) for a more detailed presentation), but we sketch the process here for completeness. There are two steps. First, we argue that x can be converted into a distribution over subsets with equivalent value for the DRO problem. Second, we argue that the *optimal* x (product distribution) has value within $(1 - 1/e)$ of the optimal arbitrary distribution over subsets.

For the first step, our starting point is the swap rounding algorithm of Chekuri et al. (2010). Swap rounding is a randomized rounding algorithm which takes a vector x and returns a feasible subset S. For any single submodular function and its multilinear extension F, swap rounding guarantees $\mathbb{E}[f(S)] \geq F(x)$. In our setting, such guarantees cannot be obtained for a single S since we want to simultaneously match the value of x with respect to n submodular functions $f_1, \dots, f_n$. However, swap rounding obeys a desirable concentration property which allows us to form a distribution $\mathcal{D}$ by running swap rounding independently several times and returning the empirical distribution over the outputs. Provided that we take sufficiently many samples, $\mathcal{D}$ is guaranteed to satisfy $\mathbb{E}_{S\sim\mathcal{D}}[f_i(S)] \geq F_i(x) - \epsilon$ for all $i = 1...n$ with high probability. Specifically, Wilder (2018) show that it suffices to draw $O\left(\frac{\log\frac{n}{\delta}}{\epsilon^3}\right)$ sets via swap rounding in order for this guarantee to hold with probability $1-\delta$.

The other piece of Lemma 3.1 relates the optimal value for Problem (7) (optimizing over product distributions) to the optimal value for the complete DRO problem (optimizing over arbitrary distributions). These values are easily shown to be within $(1-1/e)$ of each other by applying the correlation gap result of Agrawal et al. (2010). For any product distribution p over subsets, let $\mathrm{marg}(p)$ denote the set of (potentially correlated) distributions with the same marginals as p. This result shows that for any submodular function f,

$$\max_{p\text{: a product distribution}} \max_{q\in\mathrm{marg}(p)} \frac{\mathbb{E}_{S\sim q}[f(S)]}{\mathbb{E}_{S\sim p}[f(S)]} \leq \frac{e}{e-1}$$

and now Lemma 3.1 follows by applying the correlation gap bound to each of the f_i.