Gradient Methods for Submodular Maximization

Hamed Hassani*† Mahdi Soltanolkotabi*‡ Amin Karbasi§

May, 2017; Revised December 2017

Abstract

In this paper, we study the problem of maximizing continuous submodular functions that naturally arise in many learning applications such as those involving utility functions in active learning and sensing, matrix approximations and network inference. Despite the apparent lack of convexity in such functions, we prove that stochastic projected gradient methods can provide strong approximation guarantees for maximizing continuous submodular functions with convex constraints. More specifically, we prove that for monotone continuous DR-submodular functions, all fixed points of projected gradient ascent provide a factor $1/2$ approximation to the global maxima. We also study stochastic gradient and mirror methods and show that after $\mathcal{O}(1/\epsilon^2)$ iterations these methods reach solutions which achieve in expectation objective values exceeding $(\frac{\mathrm{OPT}}{2} - \epsilon)$. An immediate application of our results is to maximize submodular functions that are defined stochastically, i.e. the submodular function is defined as an expectation over a family of submodular functions with an unknown distribution. We will show how stochastic gradient methods are naturally well-suited for this setting, leading to a factor $1/2$ approximation when the function is monotone. In particular, it allows us to approximately maximize discrete, monotone submodular optimization problems via projected gradient ascent on a continuous relaxation, directly connecting the discrete and continuous domains. Finally, experiments on real data demonstrate that our projected gradient methods consistently achieve the best utility compared to other continuous baselines while remaining competitive in terms of computational effort.

1 Introduction

Submodular set functions exhibit a natural diminishing returns property, resembling concave functions in continuous domains. At the same time, they can be minimized exactly in polynomial time (while can only be maximized approximately), which makes them similar to convex functions. They have found numerous applications in machine learning, including viral marketing [1], dictionary learning [2] network monitoring [3, 4], sensor placement [5], product recommendation [6, 7], document and corpus summarization [8] data summarization [9], crowd teaching [10, 11], and probabilistic models [12, 13]. However, submodularity is in general a property that goes beyond set functions and can be defined for continuous functions. In this paper, we consider the following *stochastic* continuous submodular optimization problem:

$$\max_{x \in \mathcal{K}} F(x) \doteq \mathbb{E}_{\theta \sim \mathcal{D}}[F_\theta(x)], \tag{1.1}$$

*H. Hassani and M. Soltanolkotabi contributed equally.

†Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA

‡Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA

§Departments of Electrical Engineering and Computer Science, Yale University, New Haven, CT

where $\mathcal{K}$ is a bounded convex body, $\mathcal{D}$ is generally an *unknown* distribution, and F_θ's are continuous submodular functions for every $\boldsymbol{\theta} \in \mathcal{D}$. We also denote the optimum value as $\text{OPT} \triangleq \max_{\boldsymbol{x}\in\mathcal{K}} F(\boldsymbol{x})$. We note that the function $F(x)$ is itself also continuous submodular as a non-negative combination of submodular functions are still submodular [14]. The formulation (1.1) covers popular instances of submodular optimization. For instance, when $\mathcal{D}$ puts all the probability mass on a single function, (1.1) reduces to *deterministic* continuous submodular optimization. Another common objective is the *finite-sum* continuous submodular optimization where $\mathcal{D}$ is uniformly distributed over m instances, i.e., $F(x) \doteq \frac{1}{m}\sum_{\theta=1}^{m} F_\theta(x)$.

A natural approach to solving problems of the form (1.1) is to use projected stochastic methods. As we shall see in Section 5, these local search heuristics are surprisingly effective. However, the reasons for this empirical success is completely unclear. The main challenge is that maximizing F corresponds to a nonconvex optimization problem (as the function F is not concave), and a priori it is not clear why gradient methods should yield a reliable solution. This leads us to the main challenge of this paper

> Do projected gradient methods lead to *provably good solutions* for continuous submodular maximization with general convex constraints?

We answer the above question in the affirmative, proving that projected gradient methods produce a competitive solution with respect to the optimum. More specifically, given a general bounded convex body $\mathcal{K}$ and a continuous function F that is monotone, smooth, and (weakly) DR-submodular we show that

- All stationary points of a DR-submodular function F over $\mathcal{K}$ provide a $1/2$ approximation to the global maximum. Thus, projected gradient methods with sufficiently small step sizes (a.k.a. gradient flows) always lead to a solutions with $1/2$ approximation guarantees.
- Projected gradient ascent after $O\left(\frac{L_2}{\epsilon}\right)$ iterations produces a solution with objective value larger than $(\text{OPT}/2 - \epsilon)$. When calculating the gradient is difficult but an unbiased estimate can be easily obtained, stochastic projected gradient ascent finds a solution with objective value exceeding $(\text{OPT}/2 - \epsilon)$, after $O\left(\frac{L_2}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right)$ iterations. Here, L_2 is the smoothness of the continuous submodular function measured in the ℓ_2-norm, σ^2 is the variance of the stochastic gradient with respect to the true gradient and OPT is the function value at the global optimum.
- Projected mirror ascent after $O\left(\frac{L_*}{\epsilon}\right)$ iterations produces a solution with objective value larger than$(\text{OPT}/2 - \epsilon)$. Similarly, stochastic projected mirror ascent finds a solution with objective value exceeding $(\text{OPT}/2 - \epsilon)$, after $O\left(\frac{L_*}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right)$ iterations. Crucially, L_* indicates the smoothness of the continuous submodular function measured in any norm (e.g., ℓ_1) that can be substantially smaller than the ℓ_2 norm.
- More generally, for weakly continuous DR-submodular functions with parameter γ (define in (2.6)) we prove the above results with $\gamma^2/(1+\gamma^2)$ approximation guarantee.

Our result have some important implications. First, they show that projected gradient methods are an efficient way of maximizing the multilinear extension of (weakly) submodular set functions for any submodularity ratio γ (note that $\gamma = 1$ corresponds to submodular functions) [2]. Second, in contrast to conditional gradient methods for submodular maximization that should always start

from the origin [15, 16], projected gradient methods can start from any initial point in the constraint set $\mathcal{K}$ and still produce a competitive solution. Third, such conditional gradient methods, when applied to the stochastic setting (with a fixed batch size), perform poorly and can produce arbitrarily bad solutions when applied to continuous submodular functions (see Appendix B for an example and further discussion on why conditional gradient methods do not easily admit stochastic variants). In contrast, stochastic projected gradient methods are stable by design and provide a solution with 1/2 guarantee in expectation. Finally, our work provides a unifying approach for solving the *stochastic submodular maximization problem* [17]

$$f(S) \doteq \mathbb{E}_{\theta\sim\mathcal{D}}[f_\theta(S)], \tag{1.2}$$

where the functions $f_\theta : 2^V \to \mathbb{R}_+$ are submodular set functions defined over the ground set V. Such objective functions naturally arise in many data summarization applications [18] and have been recently introduced and studied in [17]. Since $\mathcal{D}$ is unknown, problem (1.2) cannot be directly solved. Instead, [17] showed that in the case of coverage functions, it is possible to efficiently maximize f by lifting the problem to the continuous domain and using stochastic gradient methods on a continuous relaxation to reach a solution that is within a factor $(1 - 1/e)$ of the optimum. In contrast, our work provides a general recipe with 1/2 approximation guarantee for problem (1.2) in which the f_θ's can be any monotone submodular function.

2 Continuous submodular maximization

A set function $f : 2^V \to \mathbb{R}_+$, defined on the ground set V, is called submodular if for all subsets $A, B \subseteq V$, we have

$$f(A) + f(B) \geq f(A \cap B) + f(A \cup B).$$

Even though submodularity is mostly considered on discrete domains, the notion can be naturally extended to arbitrary lattices [19]. To this aim, let us consider a subset of $\mathbb{R}^n_+$ of the form $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$ where each $\mathcal{X}_i$ is a compact subset of $\mathbb{R}_+$. A function $F : \mathcal{X} \to \mathbb{R}_+$ is *submodular* if for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{X}$, we have

$$F(\boldsymbol{x}) + F(\boldsymbol{y}) \geq F(\boldsymbol{x} \vee \boldsymbol{y}) + F(x \wedge \boldsymbol{y}), \tag{2.1}$$

where $\boldsymbol{x} \vee \boldsymbol{y} \doteq \max(\boldsymbol{x}, \boldsymbol{y})$ (component-wise) and $\boldsymbol{x} \wedge \boldsymbol{y} \doteq \min(\boldsymbol{x}, \boldsymbol{y})$ (component-wise). A submodular function is monotone if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ such that $\boldsymbol{x} \leq \boldsymbol{y}$, we have $F(\boldsymbol{x}) \leq F(\boldsymbol{y})$ (here, by $\boldsymbol{x} \leq \boldsymbol{y}$ we mean that every element of $\boldsymbol{x}$ is less than that of $\boldsymbol{y}$). Like set functions, we can define submodularity in an equivalent way, reminiscent of diminishing returns, as follows [14]: the function F is submodular if for any $\boldsymbol{x} \in \mathcal{X}$ and two distinct basis vectors $\mathrm{e}_i, \mathrm{e}_j \in \mathbb{R}^n$ and two non-negative real numbers $z_i, z_j \in \mathbb{R}_+$, such that $\boldsymbol{x}_i + z_i \in \mathcal{X}_i$ and $\boldsymbol{x}_j + z_j \in \mathcal{X}_j$, then

$$F(\boldsymbol{x} + z_i \mathrm{e}_i) + F(\boldsymbol{x} + z_j \mathrm{e}_j) \geq F(\boldsymbol{x}) + F(\boldsymbol{x} + z_i \mathrm{e}_i + z_j \mathrm{e}_j). \tag{2.2}$$

Clearly, the above definition includes submodularity over a set (by restricting $\mathcal{X}_i$'s to $\{0, 1\}$) or over an integer lattice (by restricting $\mathcal{X}_i$'s to $\mathbb{Z}_+$) as special cases. However, in the remaining of this paper we consider *continuous* submodular functions defined on product of sub-intervals of $\mathbb{R}_+$. When twice differentiable, F is submodular if and only if all cross-second-derivatives are non-positive [14], i.e.,

$$\forall i \neq j, \forall \boldsymbol{x} \in \mathcal{X}, \quad \frac{\partial^2 F(\boldsymbol{x})}{\partial x_i \partial x_j} \leq 0. \tag{2.3}$$

The above expression makes it clear that continuous submodular functions are not convex nor concave in general, as concavity (convexity) implies that $\nabla^2 F \preceq 0$ (resp.$\nabla^2 F \succeq 0$). Indeed, we can have functions that are both submodular and convex/concave. For instance, for a concave function g and non-negative weights $\lambda_i \geq 0$, the function $F(\boldsymbol{x}) = g(\sum_{i=1}^{n} \lambda_i x_i)$ is submodular and concave. Trivially, affine functions are submodular, concave, and convex. A proper subclass of submodular functions are called *DR-submodular* [16, 20] if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ such that $\boldsymbol{x} \leq \boldsymbol{y}$ and any standard basis vector $\mathrm{e}_i \in \mathbb{R}^n$ and a non-negative number $z \in \mathbb{R}_+$ such that $z\mathrm{e}_i + \boldsymbol{x} \in \mathcal{X}$ and $z\mathrm{e}_i + \boldsymbol{y} \in \mathcal{X}$, then,

$$F(z\mathrm{e}_i + \boldsymbol{x}) - F(\boldsymbol{x}) \geq F(z\mathrm{e}_i + \boldsymbol{y}) - F(\boldsymbol{y}). \tag{2.4}$$

One can easily verify that for a differentiable DR-submodular functions the gradient is an antitone mapping, i.e., for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ such that $\boldsymbol{x} \leq \boldsymbol{y}$ we have $\nabla F(\boldsymbol{x}) \geq \nabla F(\boldsymbol{y})$ [16]. When twice differentiable, DR-submodularity is equivalent to

$$\forall i \ \& \ j, \forall \boldsymbol{x} \in \mathcal{X}, \quad \frac{\partial^2 F(\boldsymbol{x})}{\partial x_i \partial x_j} \leq 0. \tag{2.5}$$

The above twice differentiable functions are sometimes called *smooth* submodular functions in the literature [21]. However, in this paper, we say a differentiable submodular function F is *L-smooth* w.r.t a norm $\|\cdot\|$ (and its dual norm $\|\cdot\|_*$) if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ we have

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\|_* \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

Here, $\|\cdot\|_*$ is the *dual norm* of $\|\cdot\|$ defined as $\|\boldsymbol{g}\|_* = \sup_{\boldsymbol{x} \in \mathbb{R}^n:\ \|\boldsymbol{x}\| \leq 1} \boldsymbol{g}^T \boldsymbol{x}$. When the function is smooth w.r.t the ℓ_2-norm we use L_2 (note that the ℓ_2 norm is self-dual). We say that a function is *weakly DR-submodular* with parameter γ if

$$\gamma = \inf_{\substack{x, y \in \mathcal{X} \\ \boldsymbol{x} \leq \boldsymbol{y}}} \inf_{i \in [n]} \frac{[\nabla F(\boldsymbol{x})]_i}{[\nabla F(\boldsymbol{y})]_i}. \tag{2.6}$$

See [22] for related definitions. Clearly, for a differentiable DR-submodular function we have $\gamma = 1$. An important example of a DR-submodular function is the multilinear extension [15] $F : [0,1]^n \to \mathbb{R}$ of a discrete submodular function f, namely,

$$F(\boldsymbol{x}) = \sum_{S \subseteq V} \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j) f(S).$$

We note that for set functions, DR-submodularity (i.e., Eq. 2.4) and submodularity (i.e., Eq. 2.1) are equivalent. However, this is not true for the general submodular functions defined on integer lattices or product of sub-intervals [16, 20].

The focus of this paper is on continuous submodular maximization defined in Problem (1.1). More specifically, we assume that $\mathcal{K} \subset \mathcal{X}$ is a a general bounded convex set (not necessarily down-closed as considered in [16]) with diameter R. Moreover, we consider F_θ's to be monotone (weakly) DR-submodular functions with parameter γ.

3 Background and related work

Submodular set functions [23, 19] originated in combinatorial optimization and operations research, but they have recently attracted significant interest in machine learning. Even though they are

usually considered over discrete domains, their optimization is inherently related to continuous optimization methods. In particular, Lovasz [24] showed that Lovasz extension is convex if and only if the corresponding set function is submodular. Moreover, minimizing a submodular set-function is equivalent to minimizing the Lovasz extension.[1] This idea has been recently extended to minimization of strict continuous submodular functions (i.e., cross-order derivatives in (2.3) are strictly negative) [14]. Similarly, approximate submodular maximization is linked to a different continuous extension known as multilinear extension [26]. Multilinear extension (which is an example of DR-submodular functions studied in this paper) is not concave nor convex in general. However, a variant of conditional gradient method, called *continuous greedy*, can be used to approximately maximize them. Recently, Chekuri et al [21] proposed an interesting multiplicative weight update algorithm that achieves $(1-1/e-\epsilon)$ approximation guarantee after $\tilde{O}(n^2/\epsilon^2)$ steps for twice differentiable monotone DR-submodular functions (they are also called smooth submodular functions) subject to a polytope constraint. Similarly, Bian et al [16] proved that a conditional gradient method, similar to the continuous greedy algorithm, achieves $(1-1/e-\epsilon)$ approximation guarantee after $O(L_2/\epsilon)$ iterations for maximizing a monotone DR-submodular functions subject to special convex constraints called *down-closed* convex bodies. A few remarks are in order. First, the proposed conditional gradient methods cannot handle the general stochastic setting we consider in Problem (1.1) (in fact, projection is the key). Second, there is no near-optimality guarantee if conditional gradient methods do not start from the origin. More precisely, for the continuous greedy algorithm it is necessary to start from the $\mathbf{0}$ vector (to be able to remain in the convex constraint set at each iteration). Furthermore, the $\mathbf{0}$ vector must be a feasible point of the constraint set. Otherwise, the iterates of the algorithm may fall out of the convex constraint set leading to an infeasible final solution. Third, due to the starting point requirement, they can only handle special convex constraints, called down-closed. Finally, the dependency on L_2 is very subomptimal as it can be as large as the dimension n (e.g., for the multilinear extensions of some submodular set functions, see Appendix C). Our work resolves all of these issues by showing that projected gradient methods can also approximately maximize monotone DR-submodular functions subject to general convex constraints, albeit, with a lower $1/2$ approximation guarantee.

Generalization of submodular set functions has lately received a lot of attention. For instance, a line of recent work considered DR-submodular function maximization over an integer lattice [27, 28, 20]. Interestingly, Ene and Nguyen [29] provided an efficient reduction from an integer-lattice DR-submodular to a submodular set function, thus suggesting a simple way to solve integer-lattice DR-submodular maximization. Note that such reductions cannot be applied to the optimization problem (1.1) as expressing general convex body constraints may require solving a continuous optimization problem.

4 Algorithms and main results

In this section we discuss our algorithms together with the corresponding theoretical guarantees. In what follows, we assume that F is a weakly DR-submodular function with parameter γ.

4.1 Characterizing the quality of stationary points

We begin with the definition of a stationary point.

[1] The idea of using stochastic methods for submodular minimization has recently been used in [25].

Definition 4.1 *A vector $\boldsymbol{x} \in \mathcal{K}$ is called a stationary point of a function $F : \mathcal{X} \to \mathbb{R}_+$ over the set $\mathcal{K} \subset \mathcal{X}$ if $\max_{\boldsymbol{y} \in \mathcal{K}} \langle \nabla F(x), \boldsymbol{y} - \boldsymbol{x} \rangle \leq 0$.*

Stationary points are of interest because they characterize the fixed points of the Gradient Ascent (GA) method. Furthermore, (projected) gradient ascent with a sufficiently small step size is known to converge to a stationary point for smooth functions [30]. To gain some intuition regarding this connection, let us consider the GA procedure. Roughly speaking, at any iteration t of the GA procedure, the value of F increases (to the first order) by $\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle$. Hence, the progress at time t is at most $\max_{\boldsymbol{y} \in \mathcal{K}} \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{y} - \boldsymbol{x}_t \rangle$. If at any time t we have $\max_{\boldsymbol{y} \in \mathcal{K}} \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{y} - \boldsymbol{x}_t \rangle \leq 0$, then the GA procedure will not make any progress and it will be stuck once it falls into a stationary point.

The next natural question is how small can the value of F be at a stationary point compared to the global maximum? The following lemma (which is an extension of Lemma 3.2 in [31]) relates the value of F at a stationary point to OPT.

Theorem 4.2 *Let $F : \mathcal{X} \to \mathbb{R}_+$ be monotone and weakly DR-submodular with parameter γ and assume $\mathcal{K} \subseteq \mathcal{X}$ is a convex set. Then,*

(i) If $\boldsymbol{x}$ is a stationary point of F in $\mathcal{K}$, then $F(\boldsymbol{x}) \geq \frac{\gamma^2}{1+\gamma^2}\mathrm{OPT}$.

(ii) Furthermore, if F is L-smooth, gradient ascent with a step size smaller than $1/L$ will converge to a stationary point.

The theorem above guarantees that all fixed points of the GA method yield a solution whose function value is at least $\frac{\gamma^2}{1+\gamma^2}\mathrm{OPT}$. Thus, all fixed point of GA provide a factor $\frac{\gamma^2}{1+\gamma^2}$ approximation ratio. The particular case of $\gamma = 1$, i.e., when F is DR-submodular, asserts that at any stationary point F is at least OPT/2. This lower bound is in fact tight. In Appendix A we provide a simple instance of a differentiable DR-Submodular function that attains OPT/2 at a stationary point that is also a *local maximum*.

We would like to note that our result on the quality of stationary points (i.e., first part of Theorem 4.2 above) can be viewed as a simple extension of the results in [31]. In particular, the special case of $\gamma = 1$ follows directly from [26, Lemma 3.2]. See Section 7.3 for complete detail on how this lemma is used in our proofs. However, we note that the main focus of this paper is whether such a stationary point can be found efficiently using stochastic schemes that do not require exact evaluations of gradients. This is the subject of the next section.

4.2 (Stochastic) gradient methods

We now discuss our first algorithmic approach. For simplicity we focus our exposition on the DR submodular case, i.e., $\gamma = 1$, and discuss how this extends to the more general case in the proofs (Section 7.4). A simple approach to maximizing DR submodular functions is to use the (projected) Gradient Ascent (GA) method. Starting from an initial estimate $\boldsymbol{x}_1 \in \mathcal{K}$ obeying the constraints, GA iteratively applies the following update

$$\boldsymbol{x}_{t+1} = \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{x}_t + \mu_t \nabla F(\boldsymbol{x}_t) \right). \tag{4.1}$$

Here, μ_t is the learning rate and $\mathcal{P}_{\mathcal{K}}(\boldsymbol{v})$ denotes the Euclidean projection of $\boldsymbol{v}$ onto the set $\mathcal{K}$. However, in many problems of practical interest we do not have direct access to the gradient of

Algorithm 1 (Stochastic) Gradient Method for Maximizing $F(x)$ over a convex set $\mathcal{K}$

Parameters: Integer $T > 0$ and scalars $\eta_t > 0$, $t \in [T]$
Initialize: $\boldsymbol{x}_1 \in \mathcal{K}$
for $t = 1$ **to** T **do**
  $\boldsymbol{y}_{t+1} \leftarrow \boldsymbol{x}_t + \eta_t \boldsymbol{g}_t$,
    where $\boldsymbol{g}_t$ is a random vector s.t. $\mathbb{E}[\boldsymbol{g}_t | \boldsymbol{x}_t] = \nabla F(\boldsymbol{x}_t)$
  $\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x} \in \mathcal{K}} \|\boldsymbol{x} - \boldsymbol{y}_{t+1}\|_2$
end for
Pick τ uniformly at random from $\{1, 2, \ldots, T\}$.
Output $\boldsymbol{x}_\tau$

F. In these cases it is natural to use a stochastic estimate of the gradient in lieu of the actual gradient. This leads to the Stochastic Gradient Method (SGM). Starting from an initial estimate $\boldsymbol{x}_0 \in \mathcal{K}$ obeying the constraints, SGM iteratively applies the following updates

$$\boldsymbol{x}_{t+1} = \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{x}_t + \mu_t \boldsymbol{g}_t \right). \tag{4.2}$$

Specifically, at every iteration t, the current iterate $\boldsymbol{x}_t$ is updated by adding $\mu_t \boldsymbol{g}_t$, where $\boldsymbol{g}_t$ is an unbiased estimate of the gradient $\nabla F(\boldsymbol{x}_t)$ and μ_t is the learning rate. The result is then projected onto the set $\mathcal{K}$. We note that when $\boldsymbol{g}_t = \nabla F(\boldsymbol{x}_t)$, i.e., when there is no randomness in the updates, then the SGM updates (4.2) reduce to the GA updates (4.1). We detail the SGM method in Algorithm 1.

As we shall see in our experiments detained in Section 5, the SGM method is surprisingly effective for maximizing monotone DR-submodular functions. However, the reasons for this empirical success was previously unclear. The main challenge is that maximizing F corresponds to a nonconvex optimization problem (as the function F is not concave), and a priori it is not clear why gradient methods should yield a reliable solution. Thus, studying gradient methods for such nonconvex problems poses new challenges:

> Do (stochastic) gradient methods converge to a stationary point? If so, how fast? How good of an approximation of the global optima is a stationary point?

The next theorem addresses some of these challenges. To be able to state this theorem let us recall the standard definition of smoothness. We say that a continuously differentiable function F is L-smooth (in Euclidean norm) if the gradient ∇F is L-Lipschitz, that is $\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|_{\ell_2} \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}$. We also defined the diameter (in Euclidean norm) as $R^2 = \sup_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}} \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}^2$. We now have all the elements in place to state our first theorem.

Theorem 4.3 (Stochastic Gradient Method) *Let us assume that F is L-smooth w.r.t. the Euclidean norm $\|\cdot\|_{\ell_2}$, monotone and DR-submodular. Furthermore, assume that we have access to a stochastic oracle $\boldsymbol{g}_t$ obeying*

$$\mathbb{E}[\boldsymbol{g}_t] = \nabla F(\boldsymbol{x}_t) \quad \text{and} \quad \mathbb{E}\left[\|\boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t)\|_{\ell_2}^2 \right] \leq \sigma^2.$$

We run stochastic gradient updates of the form (4.2) with $\mu_t = \frac{1}{L + \frac{\sigma}{R}\sqrt{t}}$. Let τ be a random variable taking values in $\{1, 2, \ldots, T\}$ with equal probability. Then,

$$\mathbb{E}[F(\boldsymbol{x}_\tau)] \geq \frac{\mathrm{OPT}}{2} - \left(\frac{R^2 L + OPT}{2T} + \frac{R\sigma}{\sqrt{T}} \right). \tag{4.3}$$

Remark 4.4 *We would like to note that if we pick τ to be a random variable taking values in $\{2, \dots, T-1\}$ with probability $\frac{1}{(T-1)}$ and 1 and T each with probability $\frac{1}{2(T-1)}$ then*

$$\mathbb{E}[F(\boldsymbol{x}_\tau)] \geq \frac{\mathrm{OPT}}{2} - \left(\frac{R^2 L}{2T} + \frac{R\sigma}{\sqrt{T}}\right).$$

The above results roughly state that $T = \mathcal{O}\left(\frac{R^2 L}{\epsilon} + \frac{R^2\sigma^2}{\epsilon^2}\right)$ iterations of the stochastic gradient method from any initial point, yields a solution whose objective value is at least $\frac{\mathrm{OPT}}{2} - \epsilon$. Stated differently, $T = \mathcal{O}\left(\frac{R^2 L}{\epsilon} + \frac{R^2\sigma^2}{\epsilon^2}\right)$ iterations of the stochastic gradient method provides in expectation a value that exceeds $\frac{\mathrm{OPT}}{2} - \epsilon$ approximation ratio for DR-submodular maximization. As explained in Section 4.1, it is not possible to go beyond the factor 1/2 approximation ratio using gradient ascent from an arbitrary initialization.

An important aspect of the above result is that it only requires an unbiased estimate of the gradient. This flexibility is crucial for many DR-submodular maximization problems (see, (1.1)) as in many cases calculating the function F and its derivative is not feasible. However, it is possible to provide a good un-biased estimator for these quantities.

We would like to point out that our results are similar in nature to known results about stochastic methods for convex optimization. Indeed, this result interpolates between the $\frac{1}{\sqrt{T}}$ for stochastic smooth optimization, and the $1/T$ for deterministic smooth optimization. The special case of $\sigma = 0$ which corresponds to Gradient Ascent deserves particular attention. In this case, and under the assumptions of Theorem 4.3, it is possible to show that $F(\boldsymbol{x}_T) \geq \frac{\mathrm{OPT}}{2} - \frac{R^2 L}{T}$, without the need for a randomized choice of $\tau \in [T]$.

Finally, we would like to note that while the first term in (4.4) decreases as $1/T$, the pre-factor L could be rather large in many applications. For instance, this quantity may depend on the dimension of the input n (see Section C in the Appendix). Thus, the number of iterations for reaching a desirable accuracy may be very large. Such a large computational load causes (stochastic) gradient methods infeasible in some application domains. We will overcome this deficiency in the next section by using stochastic mirror methods.

4.3 Stochastic mirror method

In the previous section we saw that when the function F and the constraint set $\mathcal{K}$ are well-behaved in the Euclidean norm (e.g., L is a constant in the ℓ_2 norm) then the total number of iterations to reach a certain accuracy is dimension-free and independent of the ambient dimension n. However, in many cases of interest, including some that arise from the multilinear extension of discrete submodular functions, the smoothness parameter scales with the ambient dimension and thus the number of iterations will be dimension dependent. This is particularly problematic for large-scale applications where the ambient dimension is very large. However, smoothness when measured in a different norm may still be dimension independent. Indeed, multilinear relaxation of discrete submodular functions have smoothness parameter in ℓ_1 norm that is bounded by their maximum singleton value (see Section C in the appendices). In this section we discuss our results for Mirror methods which are designed to adapt to smoothness in general norms. To explain the mirror descent method we need a few definitions. First, recall the definition of smoothness to arbitrary norms: a continuously differentiable function F is L-smooth with respect to a norm $\|\cdot\|$ if the gradient ∇F obeys $\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|_* \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$. Here, $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ defined as

Algorithm 2 (Stochastic) Mirror Ascent for Maximizing $F(x)$ over a convex set $\mathcal{K}$

Parameters: Integer $T > 0$ and scalars $\mu_t > 0$, $t \in [T]$
Initialize: $x_1 \in \mathcal{K}$
for $t = 1$ **to** T **do**
  $\nabla\phi(\boldsymbol{y}_{t+1}) \leftarrow \nabla\phi(\boldsymbol{x}_t) + \mu_t \boldsymbol{g}_t$ with $\boldsymbol{g}_t$ obeying $\mathbb{E}[\boldsymbol{g}_t|\boldsymbol{x}_t] = \nabla F(\boldsymbol{x}_t)$
  $\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{K}} \mathcal{D}_\Phi(\boldsymbol{x}, \boldsymbol{y}_{t+1})$ where $\mathcal{D}_\Phi$ is the Bregman divergence associated with the mirror map
end for
Pick τ uniformly at random from $\{1, 2, \ldots, T\}$.
Output x_τ

$\|\boldsymbol{g}\|_* = \sup_{\boldsymbol{x}\in\mathbb{R}^n:\, \|\boldsymbol{x}\|\le 1} \boldsymbol{g}^T\boldsymbol{x}$. We also need the definition of the mirror map and Bregman divergence (our exposition is adapted from [32]).

Definition 4.5 (mirror map) *Let $\mathcal{D} \subset \mathbb{R}^n$ be a convex open set. We say that $\Phi : \mathcal{D} \to \mathbb{R}$ is a mirror map if it satisfies the following properties:*

(a) Φ is strictly convex and differentiable.

(b) The gradient of Φ takes all possible values, that is $\nabla\Phi(\mathcal{D}) = \mathbb{R}^n$.

(c) The gradient of Φ diverges on the boundary of $\mathcal{D}$, that is $\lim_{\boldsymbol{x}\to\partial\mathcal{D}} \|\nabla\Phi(\boldsymbol{x})\| = +\infty$. We study mirror maps with $\mathcal{D} = \mathbb{R}^n_+$ equal to the positive orthant.

Definition 4.6 (Bregman Divergence and Projection) *We define the Bregman divergence associated to a mirror map ϕ as*

$$D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle\nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle.$$

We also define the projection onto a set $\mathcal{K}$ with respect to a mapping Φ via

$$\Pi^\Phi_\mathcal{K}(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}\in\mathcal{K}} \mathcal{D}_\Phi(\boldsymbol{x}, \boldsymbol{y}).$$

Finally, we define the diameter as follows

$$R^2 = \sup_{\boldsymbol{x}\in\mathcal{K}, \boldsymbol{y}\in\mathcal{K}\cap\mathbb{R}^n_{++}} \Phi(\boldsymbol{x}) - \Phi(\boldsymbol{y}).$$

We are now ready to describe the stochastic mirror method based on a mirror map Φ and let $\boldsymbol{x}_1 \in \arg\min_{\boldsymbol{x}\in\mathcal{K}} \Phi(\boldsymbol{x})$. Also, let $\boldsymbol{g}_t$ be an unbiased estimator of the gradient $\nabla f(\boldsymbol{x}_t)$. Then for $t \ge 1$, let $\boldsymbol{y}_{t+1} \in \mathbb{R}^n_+$ be such that $\nabla\Phi(\boldsymbol{y}_{t+1}) = \nabla\Phi(\boldsymbol{x}_t) - \mu_t\boldsymbol{g}_t$. Using $\boldsymbol{y}_{t+1}$ we obtain the next estimate $\boldsymbol{x}_{t+1}$ by projecting $\boldsymbol{y}_{t+1}$ onto $\mathcal{K}$ using the mirror map. We detail the Stochastic Mirror Ascent in Algorithm 2.

Theorem 4.7 (Stochastic Mirror Method) *Let Φ be a mirror map that is 1-strongly convex on $\mathcal{K}$ with respect to the norm $\|\|$. Assume that F is L-smooth with respect to the norm $\|\|$ and*

is a monotone, continuous submodular function. Furthermore, assume that we have access to a stochastic oracle $\boldsymbol{g}_t$ *obeying*

$$\mathbb{E}[\boldsymbol{g}_t] = \nabla F(\boldsymbol{x}_t) \quad \textit{and} \quad \mathbb{E}\left[\|\boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t)\|_*^2\right] \leq \sigma^2.$$

We start from $\boldsymbol{x}_1 \in \arg\min_{\boldsymbol{x}\in\mathcal{K}} \Phi(\boldsymbol{x})$ *and run the mirror ascent updates of the form*

$$\begin{aligned} \nabla\Phi(\boldsymbol{y}_{t+1}) &= \nabla\Phi(\boldsymbol{x}_t) + \mu_t \boldsymbol{g}_t, \\ \boldsymbol{x}_{t+1} &= \Pi_{\mathcal{K}}^{\Phi}\left(\boldsymbol{y}_{t+1}\right), \end{aligned}$$

with $\mu_t = \frac{1}{L + \frac{\sigma}{R}\sqrt{t}}$. *Let* τ *be a random variable taking values in* $\{1, 2, \ldots, T\}$ *with equal probability. Then,*

$$\mathbb{E}[F(\boldsymbol{x}_\tau)] \geq \frac{\mathrm{OPT}}{2} - \left(\frac{R^2 L + OPT}{2T} + \frac{R\sigma}{\sqrt{T}}\right). \tag{4.4}$$

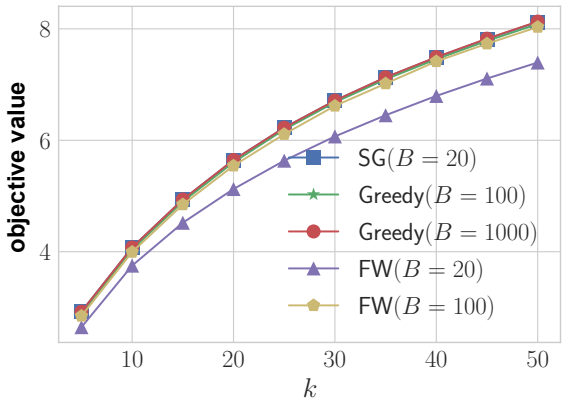
Remark 4.8 *We would like to note that if we pick* τ *to be a random variable taking values in* $\{2, \ldots, T-1\}$ *with probability* $\frac{1}{(T-1)}$ *and* 1 *and* T *each with probability* $\frac{1}{2(T-1)}$ *then*

$$\mathbb{E}[F(\boldsymbol{x}_\tau)] \geq \frac{\mathrm{OPT}}{2} - \left(\frac{R^2 L}{2T} + \frac{R\sigma}{\sqrt{T}}\right).$$

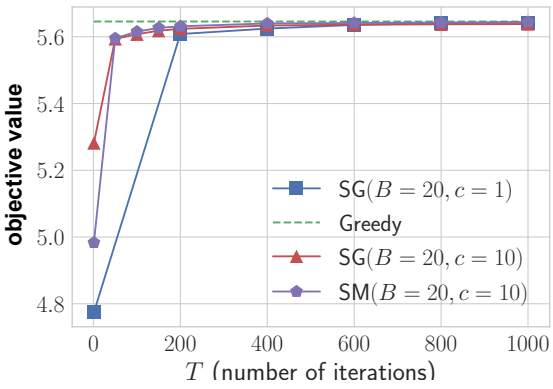
As a simple application of Theorem 4.7, let us consider submodular optimization problems that arise from maximizing submodular set functions under k-cardinality constraints. For such problems, it will be convenient to use the mirror ascent method with ℓ_1 norm on the scaled simplex $\{\boldsymbol{z} \in [0,1]^n : \sum_{i=1}^n \boldsymbol{z}_i = k\}$ with the entropy mirror map $\Phi(\boldsymbol{x}) = k\sum_{i=1}^n x(i)\log x(i)$. This is due to the fact that the smoothness parameter of the multilinear extension in the ℓ_1 norm might be much smaller than its counterpart in the ℓ_1 norm. In this case, the updates in Algorithm 2 take the form

$$\begin{aligned} [\boldsymbol{y}_{t+1}]_i &= [\boldsymbol{x}_t]_i e^{-\eta[\frac{\mu_t}{k}\boldsymbol{g}_t]_i}, \text{ for } i = 1, 2, \ldots, n, \\ \boldsymbol{x}_{t+1} &= \arg\min_{\boldsymbol{x}\in\mathcal{K}} \mathrm{KL}(\boldsymbol{x}, \boldsymbol{y}_{t+1}). \end{aligned}$$

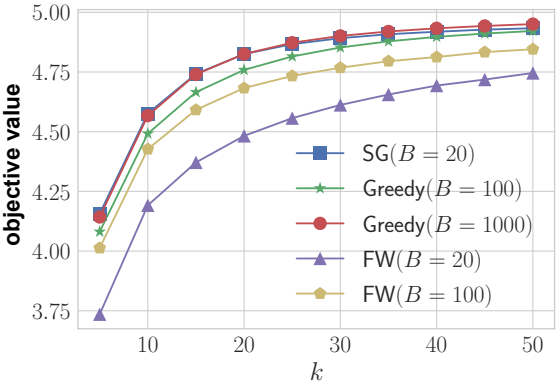
Here, $\mathrm{KL}(\boldsymbol{x}, \boldsymbol{y})$ denotes the KL divergence between the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. We also note that the corresponding projection can be be done very efficiently in $\mathcal{O}(n)$ time using standard methods described in [33, 34]. The reason for using the ℓ_1 norm with the entropy map is that the smoothness parameter L can be bounded by a constant. Indeed, the smoothness parameter of the multilinear extension of a monotone submodular function f can be bounded by the maximum marginal value of f, i.e. $L \leq m_f \triangleq \max_e f(e)$. Furthermore, R^2 can be bounded by $O(k\log n)$. Thus, using this particular mirror method, the above result roughly states that $T = \mathcal{O}\left(\frac{m_f k\log(n)}{\epsilon} + \frac{k\sigma^2\log(n)}{\epsilon^2}\right)$ iterations of the stochastic mirror method, yields a solution whose objective value is at least $\frac{\mathrm{OPT}}{2} - \epsilon$. Stated differently, $T = \mathcal{O}\left(\frac{m_f k\log(n)}{\epsilon} + \frac{k\sigma^2\log(n)}{\epsilon^2}\right)$ iterations of the stochastic mirror method provides in expectation an objective value exceeding $\frac{\mathrm{OPT}}{2} - \epsilon$ approximation ratio for DR-submodular maximization. Thus, the required number of iterations to reach a certain accuracy now depends only logarithmically on n.



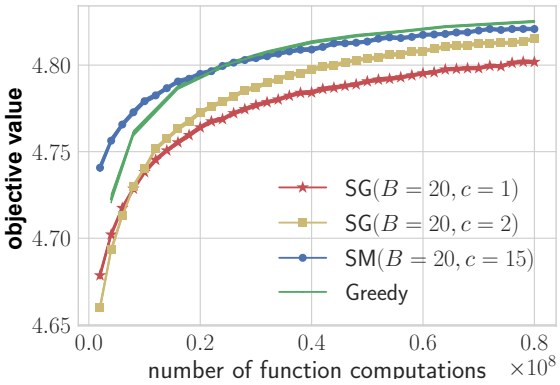
(a) Concave Over Modular



(b) Concave Over Modular



(c) Facility Location



(d) Facility Location

Figure 1: (a) shows the performance of the algorithms w.r.t. the cardinality constraint k for the concave over modular objective. Each of the continuous algorithms (i.e., SG, SM and FW) run for $T = 2000$ iterations. (b) shows the performance of the algorithms SG and SM versus the number of iterations for fixed $k = 20$ for the concave over modular objective. The green dashed line indicates the value obtained by Greedy (with $B = 1000$). Recall that the step size of SM and SG is $c/\sqrt{t}$. (c) shows the performance of the algorithms w.r.t. the cardinality constraint k for the facility location objective function. Each of the continuous algorithms (SG, SM, FW) run for $T = 2000$ iterations. (d) shows the performance of different algorithms versus the number of simple function computations (i.e. the number of f_i's evaluated during the algorithm) for the facility location objective function. For the greedy algorithm, larger number of function computations corresponds to a larger batch size. For SG and SM, larger time corresponds to larger iterations.

5 Experiments

In our experiments, we consider a movie recommendation application [18] consisting of N users and n movies. Each user i has a user-specific utility function f_i for evaluating sets of movies. The

goal is to find a set of k movies such that in expectation over users' preferences it provides the highest utility, i.e., $\max_{|S|\leq k} f(S)$, where $f(S) \doteq \mathbb{E}_{i\sim\mathcal{D}}[f_i(S)]$. This is an instance of the stochastic submodular maximization problem defined in (1.2). We consider a setting that consists of N users and consider the empirical objective function $\frac{1}{N}\sum_{j=1}^{N} f_i$. In other words, the distribution $\mathcal{D}$ is assumed to be uniform on the integers between 1 and N. We can then run the (discrete) greedy algorithm on the empirical objective function to find a good set of size k. However, as N is a large number, the greedy algorithm will require a high computational complexity. Another way of solving this problem is to evaluate the multilinear extension F_i of any sampled function f_i and solve the problem in the continuous domain as follows. Let $F(\boldsymbol{x}) = \mathbb{E}_{i\sim\mathcal{D}}[F_i(\boldsymbol{x})]$ for $x \in [0,1]^n$ and define the constraint set $\mathcal{P}_k = \{\boldsymbol{x} \in [0,1]^m : \sum_{i=1}^{n} x_i \leq k\}$. The discrete and continuous optimization formulations lead to the same optimal value [15]:

$$\max_{S:|S|\leq k} f(S) = \max_{\boldsymbol{x}\in\mathcal{P}_k} F(\boldsymbol{x}).$$

Therefore, by running the stochastic versions of projected gradient methods, we can find a solution in the continuous domain that provides a $1/2$ approximation to the optimal value. By rounding this fractional solution (for instance via randomized Pipage rounding [15]) we obtain a set whose utility is at least $1/2$ of the optimum solution set of size k. We note that randomized Pipage rounding does not need access to the value of f. We also remark that projection onto $\mathcal{P}_k$ can be done very efficiently in $O(n)$ time (see [17, 33, 34]). Therefore, such approach easily scales to big data scenarios where the size of the data set (e.g. number of users) or the number of items n (e.g. number of movies) are very large.
In our experiments, we consider the following baselines:

(i) Stochastic Gradient Ascent (SG): with the step size $\mu_t = c/\sqrt{t}$ and batch size B. The details for computing an unbiased estimation for the gradient of F are given in Appendix D.

(ii) Stochastic Mirror Ascent (SM): with the step size $\mu_t = c/\sqrt{t}$ and batch size B.

(iii) Frank-Wolfe (FW) variant of [16]: with parameter T for the total number of iterations and batch size B (we further let $\alpha = 1, \delta = 0$, see Algorithm 1 in [16] for more details).

(iv) Batch-mode Greedy (Greedy): by running greedy algorithm over the empirical objective function with B samples.

To run the experiments we use the MovieLens data set. It consists of 1 million ratings (from 1 to 5) by $n = 6041$ users for $m = 4000$ movies. Let $r_{i,j}$ denote the rating of user i for movie j (if such a rating does not exist we assign $r_{i,j}$ to 0). In our experiments, we consider two well motivated objective functions. The first one is the facility location where the valuation function by user i is defined as $f_i(S) = \max_{j\in S} r_{i,j}$. In words, the way user i evaluates a set S is by picking the highest rated movie in S. For simplicity, we also assume that the distribution $\mathcal{D}$ is uniform. Thus, the objective function is equal to

$$f_{\text{fac}}(S) = \frac{1}{N}\sum_{i=1}^{N} \max_{j\in S} r_{i,j}.$$

In our second experiment, we consider a different user-specific valuation function which is a concave function composed with a modular function, i.e., $f_i(S) = (\sum_{j\in S} r_{i,j})^{1/2}$. Again, by considering

the uniform distribution over the set of users, we obtain

$$f_{\text{con}}(S) = \frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{j\in S} r_{i,j}\Big)^{1/2}.$$

Note that the multilinear extensions of f_{fac} and f_{con} are neither concave nor convex.

Figure 1 depicts the performance of different algorithms for the two proposed objective functions. As Figures 1a and 1c show, the FW algorithm needs a much higher batch size to be comparable in performance w.r.t. to our stochastic gradient methods. Note that a smaller batch size leads to less computational effort (using the same value for B and T, the computational complexity of FW and SGA is almost the same). Figure 1b shows that after a few hundred iterations both SG and SM with $B = 20$ obtain almost the same utility as Greedy with a large batch size ($B = 1000$). Finally, Figure 1d shows the performance of the algorithms with respect to the number of times the single functions (f_i's) are evaluated. This further shows that gradient based methods have comparable complexity w.r.t. the Greedy algorithm in the discrete domain.

6 Conclusion

In this paper we studied gradient methods for submodular maximization. Despite the lack of convexity of the objective function we demonstrated that local search heuristics are effective at finding approximately optimal solutions. In particular, we showed that all fixed point of projected gradient ascent provide a factor 1/2 approximation to the global maxima. We also demonstrated that stochastic gradient and mirror methods achieve an objective value of $\text{OPT}/2 - \epsilon$ in $\mathcal{O}(\frac{1}{\epsilon^2})$ iterations. We further demonstrated the effectiveness of our methods with experiments on real data.

While in this paper we have focused on convex constraints, our framework may allow non-convex constraints as well. For instance it may be possible to combine our framework with recent results in [35, 36, 37] to deal with general nonconvex constraints. Furthermore, in some cases projection onto the constraint set may be computationally intensive or even intractable but calculating an approximate projection may be possible with significantly less effort. One of the advantages of gradient descent-based proofs is that they continue to work even when some perturbations are introduced in the updates. Therefore, we believe that our framework can deal with approximate projections and we hope to pursue this in future work.

7 Proofs

Throughout this section we use the assumption that $\mathcal{K} \subseteq \mathcal{X} \subset \mathbb{R}^n_+$. We first prove Theorems 4.2 and 4.7. We then show in Section 7.3 how the proof of Theorem 4.3 follows from the proof of Theorem 4.7.

7.1 Proofs for quality of stationary points (Proof of Theorem 4.2)

Let us begin with part (i) of the theorem. Let $F : \mathcal{X} \to \mathbb{R}_+$ be a weakly DR-submodular and monotone function. By (2.6) for any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ we have

$$\nabla F(\boldsymbol{x}) \geq \gamma \nabla F(\boldsymbol{y}) \quad \text{for all} \quad \boldsymbol{x} \preceq \boldsymbol{y}. \tag{7.1}$$

To prove part (i) we first prove that for any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$, we have

$$F(\boldsymbol{y}) - \left(1 + \frac{1}{\gamma^2}\right) F(\boldsymbol{x}) \leq \frac{1}{\gamma} \langle \nabla F(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle. \tag{7.2}$$

To this aim note that for all $\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}$ s.t. $\boldsymbol{x} \leq \boldsymbol{z}$, by using (7.1), we have

$$\begin{aligned} F(\boldsymbol{z}) - F(\boldsymbol{x}) =& \int_0^1 \langle \boldsymbol{z} - \boldsymbol{x}, \nabla F\left(\boldsymbol{x} + t(\boldsymbol{z} - \boldsymbol{x})\right) \rangle dt, \\ \leq& \frac{1}{\gamma} \int_0^1 \langle \boldsymbol{z} - \boldsymbol{x}, \nabla F(\boldsymbol{x}) \rangle dt, \\ =& \frac{1}{\gamma} \langle \boldsymbol{z} - \boldsymbol{x}, \nabla F(\boldsymbol{x}) \rangle. \end{aligned} \tag{7.3}$$

Similarly, note that using (7.1)

$$\begin{aligned} F(\boldsymbol{z}) - F(\boldsymbol{x}) =& \int_0^1 \langle \boldsymbol{z} - \boldsymbol{x}, \nabla F\left(\boldsymbol{x} + t(\boldsymbol{z} - \boldsymbol{x})\right) \rangle dt, \\ \geq& \gamma \int_0^1 \langle \boldsymbol{z} - \boldsymbol{x}, \nabla F(\boldsymbol{z}) \rangle dt, \\ =& \gamma \langle \boldsymbol{z} - \boldsymbol{x}, \nabla F(\boldsymbol{z}) \rangle. \end{aligned} \tag{7.4}$$

Now, from (7.3) we deduce that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$:

$$F(\boldsymbol{x} \vee \boldsymbol{y}) - F(\boldsymbol{x}) \leq \frac{1}{\gamma} \langle \boldsymbol{x} \vee \boldsymbol{y} - \boldsymbol{x}, \nabla F(\boldsymbol{x}) \rangle,$$

and from (7.4):

$$F(\boldsymbol{x}) - F(\boldsymbol{x} \wedge \boldsymbol{y}) \geq \gamma \langle \boldsymbol{x} - \boldsymbol{x} \wedge \boldsymbol{y}, \nabla F(\boldsymbol{x}) \rangle,$$

From these two inequalities we immediately obtain

$$F(\boldsymbol{x} \vee \boldsymbol{y}) - (1 + \frac{1}{\gamma^2}) F(\boldsymbol{x}) + \frac{1}{\gamma^2} F(\boldsymbol{x} \wedge \boldsymbol{y}) \leq \frac{1}{\gamma} \langle \boldsymbol{x} \wedge \boldsymbol{y} + \boldsymbol{x} \vee \boldsymbol{y} - 2\boldsymbol{x}, \nabla F(\boldsymbol{x}) \rangle, \tag{7.5}$$

and we obtain (7.2) by noting that $F(\boldsymbol{x} \wedge \boldsymbol{y}) \geq 0$ and $\boldsymbol{x} \wedge \boldsymbol{y} + \boldsymbol{x} \vee \boldsymbol{y} = \boldsymbol{x} + \boldsymbol{y}$.

Part (i) of the theorem follows from (7.2) by letting $\boldsymbol{x}$ to be a stationary point and $\boldsymbol{y} = \boldsymbol{x}^* := \arg\max_{\mathcal{K}} F(\boldsymbol{y})$.

To prove part (ii) note that by the smoothness of the function (more specifically the quadratic upper bound) we have

$$F(\boldsymbol{x}_{t+1}) \geq F(\boldsymbol{x}_t) + \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle - \frac{L}{2} \left\| \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \right\|_{\ell_2}^2.$$

Now note that $\boldsymbol{x}_{t+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{x}_t + \mu_t \nabla F(\boldsymbol{x}_t)\right)$ and thus using the properties of convex projections we have

$$\langle \boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \boldsymbol{x}_{t+1} - (\boldsymbol{x}_t + \mu_t \nabla F(\boldsymbol{x}_t)) \rangle \leq 0 \quad \Rightarrow \quad \left\| \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \right\|_{\ell_2}^2 \leq \mu_t \langle \boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \nabla F(\boldsymbol{x}_t) \rangle.$$

Plugging this into the latter inequality we conclude that for $\mu_t \leq \frac{1}{L}$

$$F(\boldsymbol{x}_{t+1}) \geq F(\boldsymbol{x}_t) + \left(\frac{1}{\mu_t} - \frac{L}{2}\right) \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_{\ell_2}^2 \geq F(\boldsymbol{x}_t) + \frac{L}{2} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_{\ell_2}^2 .$$

Summing both sides we conclude that

$$\sum_{t=1}^{\infty} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_{\ell_2}^2 ,$$

is bounded. This in turn implies that $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_{\ell_2}$ goes to zero. To proceed we define the convex set $\mathcal{D}_t = \mathcal{K} - \{\boldsymbol{x}_t\}$. Note that

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_{\ell_2} = \|\mathcal{P}_{\mathcal{K}} (\boldsymbol{x}_t + \mu_t \nabla F(\boldsymbol{x}_t)) - \boldsymbol{x}_t\|_{\ell_2} = \left\|\mathcal{P}_{\mathcal{K}-\{\boldsymbol{x}_t\}} (\mu_t \nabla F(\boldsymbol{x}_t))\right\|_{\ell_2} = \|\mathcal{P}_{\mathcal{D}_t} (\mu_t \nabla F(\boldsymbol{x}_t))\|_{\ell_2} .$$

Thus,

$$\lim_{t \to \infty} \mathcal{P}_{\mathcal{D}_t} (\mu_t \nabla F(\boldsymbol{x}_t)) = 0. \tag{7.6}$$

To use the latter to show convergence to a stationary point note that for any $\boldsymbol{z} \in \mathcal{D}_t$ by the obtuse angle theorem for convex sets (e.g. see [32, Lemma 3.1])

$$\langle \mathcal{P}_{\mathcal{D}_t} (\mu_t \nabla F(\boldsymbol{x}_t)) - \boldsymbol{z}, \mathcal{P}_{\mathcal{D}_t} (\mu_t \nabla F(\boldsymbol{x}_t)) - \mu_t \nabla F(\boldsymbol{x}_t) \rangle \leq 0.$$

holds for all $\boldsymbol{z} \in \mathcal{D}_t$. Note that any $\boldsymbol{z} \in \mathcal{D}_t$ is of the form $\boldsymbol{y} - \boldsymbol{x}_t$ with $\boldsymbol{y} \in \mathcal{K}$. Thus the above inequality implies that for all $\boldsymbol{y} \in \mathcal{K}$

$$\langle \mathcal{P}_{\mathcal{D}_t} (\mu_t \nabla F(\boldsymbol{x}_t)) - (\boldsymbol{y} - \boldsymbol{x}_t) , \mathcal{P}_{\mathcal{D}_t} (\mu_t \nabla F(\boldsymbol{x}_t)) - \mu_t \nabla F(\boldsymbol{x}_t) \rangle \leq 0.$$

Now taking the limit as $t \to \infty$ and using (7.6) we conclude that for all $\boldsymbol{y} \in \mathcal{K}$

$$\lim_{t \to \infty} \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{y} - \boldsymbol{x}_t \rangle \leq 0,$$

concluding the proof.

7.2 Proof of (stochastic) mirror method (Proof of Theorem 4.7)

We begin by stating some lemmas about mirror descent together with some useful preliminary lemmas in the next section.

7.2.1 Preliminary lemmas

We begin with two lemmas about mirror descent adapted from [32].

Lemma 7.1 *Let $\boldsymbol{x} \in \mathcal{K}$ and $\boldsymbol{y} \in \mathcal{K}$, then*

$$\left\langle \nabla \Phi \left(\Pi_{\mathcal{K}}^{\Phi}(\boldsymbol{y})\right) - \nabla \Phi(\boldsymbol{y}), \Pi_{\mathcal{K}}^{\Phi}(\boldsymbol{y}) - \boldsymbol{x} \right\rangle \leq 0.$$

Also we need the following well-known identity about Bregman divergences which will be useful several times in our proofs.

Lemma 7.2

$$\langle \nabla\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{z}\rangle = D_\Phi(\boldsymbol{x},\boldsymbol{y}) + D_\Phi(\boldsymbol{z},\boldsymbol{x}) - D_\Phi(\boldsymbol{z},\boldsymbol{y}).$$

We next state a lemma due to Chekuri, Vondrak, and Zenkluser.

Lemma 7.3 *[31, Lemma 3.2] Assume F is a monotone and submodular function. Then, for any two points $\boldsymbol{x},\boldsymbol{y}\in\mathcal{K}$*

$$\left\langle \boldsymbol{x}-\boldsymbol{y}, \nabla F(\boldsymbol{x})\right\rangle \le 2F(\boldsymbol{x}) - F(\max(\boldsymbol{x},\boldsymbol{y})) - F(\min(\boldsymbol{x},\boldsymbol{y})).$$

Lemma 7.4 *Consider one iteration of the mirror descent update*

$$\begin{aligned}\nabla\Phi(\boldsymbol{y}) =& \nabla\Phi(\boldsymbol{x}) + \mu G(\boldsymbol{x}),\\ \boldsymbol{x}^+ =& \Pi^\Phi_{\mathcal{K}}(\boldsymbol{y}).\end{aligned}$$

Then, for all $\boldsymbol{z}\in\mathcal{K}$

$$\langle G(\boldsymbol{x}), \boldsymbol{x}^+ - \boldsymbol{z}\rangle \ge \frac{1}{\mu}\left(D_\Phi(\boldsymbol{x}^+,\boldsymbol{x}) + D_\Phi(\boldsymbol{z},\boldsymbol{x}^+) - D_\Phi(\boldsymbol{z},\boldsymbol{x})\right).$$

Proof By Lemma 7.1

$$\left\langle \nabla\Phi(\boldsymbol{x}^+) - \nabla\Phi(\boldsymbol{y}), \boldsymbol{x}^+ - \boldsymbol{z}\right\rangle \le 0.$$

Using $\nabla\Phi(\boldsymbol{y}) = \nabla\Phi(\boldsymbol{x}) + \mu G(\boldsymbol{x})$, we conclude that

$$\begin{aligned}\left\langle G(\boldsymbol{x}), \boldsymbol{x}^+ - \boldsymbol{z}\right\rangle \ge& \frac{1}{\mu}\left\langle \Phi(\boldsymbol{x}^+) - \nabla\Phi(\boldsymbol{x}), \boldsymbol{x}^+ - \boldsymbol{z}\right\rangle,\\ =& \frac{1}{\mu}\left(D_\Phi(\boldsymbol{x}^+,\boldsymbol{x}) + D_\Phi(\boldsymbol{z},\boldsymbol{x}^+) - D_\Phi(\boldsymbol{z},\boldsymbol{x})\right),\end{aligned}$$

where the last equality follows from Lemma 7.2. ■

Lemma 7.5 *Consider the setting of Theorem 4.7 and let D_Φ be the Bregman divergence corresponding to the mirror map Φ. Let η be a nonnegative scalar with μ_t obeying*

$$\mu_t \le \frac{1}{L + \frac{1}{\eta}}.$$

Then,

$$\left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{z}\right\rangle \ge \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{z},\boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{z},\boldsymbol{x}_t)\right) + F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t+1}) - \frac{\eta}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2.$$

Proof Using smoothness of the function F we have the following chain of inequalities

$$\begin{aligned}F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{x}_t) &\ge \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle - \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2\\ &= \langle \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle + \langle \nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle - \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2\\ &\overset{(a)}{\ge} \langle \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle - \frac{\eta}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2 - \frac{1}{2}\left(L + \frac{1}{\eta}\right)\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2\\ &\overset{(b)}{\ge} \langle \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle - \frac{\eta}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2 - \left(L + \frac{1}{\eta}\right) D_\Phi(\boldsymbol{x}_{t+1},\boldsymbol{x}_t)\\ &= \langle \boldsymbol{g}_t, \boldsymbol{z} - \boldsymbol{x}_t\rangle + \langle \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{z}\rangle - \frac{\eta}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2 - \left(L + \frac{1}{\eta}\right) D_\Phi(\boldsymbol{x}_{t+1},\boldsymbol{x}_t),\end{aligned}$$

where (a) follows from the fact that $\langle \boldsymbol{a}, \boldsymbol{b} \rangle \leq \frac{1}{2\eta}\|\boldsymbol{a}\|^2 + \frac{\eta}{2}\|\boldsymbol{b}\|_*^2$ by Young's inequality and (b) follows from strong convexity of the mirror map Φ. Rearranging the above inequality we arrive at the following chain of inequalities

$$\begin{aligned}
\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{z} \rangle \geq & \langle \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{z} \rangle - \frac{\eta}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2 - \left(L + \frac{1}{\eta}\right) D_\Phi(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t) + F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t+1}) \\
\overset{(a)}{\geq} & \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{z}, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{z}, \boldsymbol{x}_t)\right) - \frac{\eta}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2 + \left(\frac{1}{\mu_t} - \left(L + \frac{1}{\eta}\right)\right) D_\Phi(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t) \\
& + F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t+1}) \\
\overset{(b)}{\geq} & \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{z}, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{z}, \boldsymbol{x}_t)\right) - \frac{\eta}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2 + F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t+1}),
\end{aligned}$$

where (a) follows from Lemma 7.4 and (b) from the choice $\mu_t \leq 1/(L + 1/\eta)$. ■

Using Lemma 7.5 with $\boldsymbol{z} = \boldsymbol{x}^*$ (Global optimum) and $\eta = \eta_t$ we have

$$\left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle \geq \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) + F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t+1}) - \frac{\eta_t}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2.$$

Using $\left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle = \left\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle + \left\langle \boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle$ we conclude that

$$\begin{aligned}
\left\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle \geq & \left\langle \boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle \\
& + \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) + F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t+1}) - \frac{\eta_t}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2.
\end{aligned}$$

Using Lemma 7.3 with $\boldsymbol{y} = \boldsymbol{x}^*$ and $\boldsymbol{x} = \boldsymbol{x}_t$ in the above inequality we conclude that

$$\begin{aligned}
F(\boldsymbol{x}_t) + F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{x}^*) \geq & \left\langle \boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle \\
& + \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) - \frac{\eta_t}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2.
\end{aligned}$$

Taking expectation of both sides we arrive at

$$\begin{aligned}
\mathbb{E}\, F(\boldsymbol{x}_t) + \mathbb{E}\, F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{x}^*) \geq & \frac{1}{\mu_t}\left(\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - \mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) - \frac{\eta_t}{2}\,\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2\right], \\
\geq & \frac{1}{\mu_t}\left(\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - \mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) - \frac{\eta_t}{2}\sigma^2.
\end{aligned}$$

Summing both sides from $t = 1$ to T we conclude that

$$\begin{aligned}
\sum_{t=1}^{T}\left[\mathbb{E}\, F(\boldsymbol{x}_t) + \mathbb{E}\, F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{x}^*)\right] \geq & \sum_{t=1}^{T}\frac{1}{\mu_t}\left(\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - \mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) - \frac{\sigma^2}{2}\sum_{t=1}^{T}\eta_t \\
= & \frac{\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{T+1})}{\mu_T} - \frac{\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_1)}{\mu_1} + \sum_{t=1}^{T-1}\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1})\left(\frac{1}{\mu_t} - \frac{1}{\mu_{t+1}}\right) \\
& - \frac{\sigma^2}{2}\sum_{t=1}^{T}\eta_t \\
\overset{(a)}{\geq} & -\frac{R^2}{\mu_1} + R^2\sum_{t=1}^{T-1}\left(\frac{1}{\mu_t} - \frac{1}{\mu_{t+1}}\right) - \frac{\sigma^2}{2}\sum_{t=1}^{T}\eta_t \\
= & -\frac{R^2}{\mu_T} - \frac{\sigma^2}{2}\sum_{t=1}^{T}\eta_t.
\end{aligned}$$

Here, (a) follows from the fact that $D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t) \le R^2$. Now using $\eta_t = \frac{R}{\sigma\sqrt{t}}$ and $\mu_t = \frac{1}{L+\frac{1}{\eta_t}}$ we arrive at

$$\sum_{t=1}^{T}\left[\mathbb{E}\,F(\boldsymbol{x}_t) + \mathbb{E}\,F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{x}^*)\right] \ge -R^2\left(L + \frac{\sigma}{R}\sqrt{T}\right) - \frac{\sigma R}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$$
$$\overset{(a)}{\ge} -\left(R^2 L + 2\sigma R\sqrt{T}\right). \tag{7.7}$$

Here, (a) follows from the fact that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le 2\sqrt{T}$. Thus

$$\sum_{t=1}^{T} 2\,\mathbb{E}[F(\boldsymbol{x}_t)] - \mathrm{OPT} = \sum_{t=1}^{T}\left[\mathbb{E}\,F(\boldsymbol{x}_t) + \mathbb{E}\,F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{x}^*)\right] + \mathbb{E}[F(\boldsymbol{x}_1)] - \mathbb{E}[F(\boldsymbol{x}_{T+1})]$$
$$\ge -\left(R^2 L + 2\sigma R\sqrt{T}\right) - \mathbb{E}[F(\boldsymbol{x}_{T+1})].$$

Dividing both sides by $2T$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\boldsymbol{x}_t)] + \frac{1}{2T}\mathbb{E}[F(\boldsymbol{x}_{T+1})] \ge \frac{\mathrm{OPT}}{2} - \left(\frac{R^2 L}{T} + \frac{2\sigma R}{\sqrt{T}}\right).$$

The proof now follows from the fact that $E[F(\boldsymbol{x}_{T+1})] \le \mathrm{OPT}$. Note also that from (7.7) we have

$$\frac{1}{T}\sum_{t=2}^{T}\mathbb{E}[F(\boldsymbol{x}_t)] + \frac{1}{2T}(\mathbb{E}[F(\boldsymbol{x}_1) + \mathbb{E}[F(\boldsymbol{x}_{T+1})]) \ge \frac{\mathrm{OPT}}{2} - \left(\frac{R^2 L}{T} + \frac{2\sigma R}{\sqrt{T}}\right).$$

Therefore, a different sampling of the $\boldsymbol{x}_t$'s (i.e. choose $\boldsymbol{x}_1, \boldsymbol{x}_{T+1}$ w.p. $1/(2T)$ and the rest with probability $1/T$–call this sampling τ') results in $\mathbb{E}[F(\boldsymbol{x}_{\tau'})] \ge \frac{\mathrm{OPT}}{2} - \left(R^2 L + 2\sigma R\sqrt{T}\right)$.

7.3 Proof of (stochastic) gradient method (Proof of Theorem 4.3)

This proof is a special case of the proof of the previous section using the mapping $\Phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_{\ell_2}^2$.

7.4 Extensions to weakly submodular functions

In this section we shall show that Theorem 4.7 extends to weakly submodular functions with the new guarantee given by

$$\mathbb{E}[F(\boldsymbol{x}_\tau)] \ge \frac{\gamma^2}{1+\gamma^2}\mathrm{OPT} - \frac{\gamma}{1+\gamma^2}\left(\frac{\mathrm{R^2L + OPT}}{\mathrm{2T}} + \frac{\mathrm{R}\sigma}{\sqrt{\mathrm{T}}}\right). \tag{7.8}$$

To this aim using Lemma 7.5 with $\boldsymbol{z} = \boldsymbol{x}^*$ (Global optimum) and $\eta = \eta_t$ we have

$$\left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle \ge \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) + F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t+1}) - \frac{\eta_t}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2.$$

Using $\left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle = \left\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle + \left\langle \boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle$ we conclude that

$$\left\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle \ge \left\langle \boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle$$
$$+ \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) + F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t+1}) - \frac{\eta_t}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2.$$

Using condition (7.2) with $\boldsymbol{y} = \boldsymbol{x}^*$ and $\boldsymbol{x} = \boldsymbol{x}_t$ in the above inequality we conclude that

$$\left(\gamma + \frac{1}{\gamma} - 1\right) F(\boldsymbol{x}_t) + F(\boldsymbol{x}_{t+1}) - \gamma F(\boldsymbol{x}^*) \geq \left\langle \boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle + \frac{1}{\mu_t}\left(D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) - \frac{\eta_t}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2.$$

Taking expectation of both sides we arrive at

$$\begin{aligned}\left(\gamma + \frac{1}{\gamma} - 1\right) \mathbb{E}\, F(\boldsymbol{x}_t) + \mathbb{E}\, F(\boldsymbol{x}_{t+1}) - \gamma F(\boldsymbol{x}^*) &\geq \frac{1}{\mu_t}\left(\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - \mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) - \frac{\eta_t}{2}\,\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|_*^2\right], \\ &\geq \frac{1}{\mu_t}\left(\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - \mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) - \frac{\eta_t}{2}\sigma^2.\end{aligned}$$

Summing both sides from $t = 1$ to T we conclude that

$$\begin{aligned}\sum_{t=1}^{T}\left[\left(\gamma + \frac{1}{\gamma} - 1\right) \mathbb{E}\, F(\boldsymbol{x}_t) + \mathbb{E}\, F(\boldsymbol{x}_{t+1}) - \gamma F(\boldsymbol{x}^*)\right] &\geq \sum_{t=1}^{T} \frac{1}{\mu_t}\left(\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1}) - \mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t)\right) - \frac{\sigma^2}{2}\sum_{t=1}^{T}\eta_t \\ &= \frac{\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{T+1})}{\mu_T} - \frac{\mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_1)}{\mu_1} + \sum_{t=1}^{T-1} \mathbb{E}\, D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_{t+1})\left(\frac{1}{\mu_t} - \frac{1}{\mu_{t+1}}\right) \\ &\quad - \frac{\sigma^2}{2}\sum_{t=1}^{T}\eta_t \\ &\overset{(a)}{\geq} -\frac{R^2}{\mu_1} + R^2 \sum_{t=1}^{T-1}\left(\frac{1}{\mu_t} - \frac{1}{\mu_{t+1}}\right) - \frac{\sigma^2}{2}\sum_{t=1}^{T}\eta_t \\ &= -\frac{R^2}{\mu_T} - \frac{\sigma^2}{2}\sum_{t=1}^{T}\eta_t.\end{aligned}$$

Here, (a) follows from the fact that $D_\Phi(\boldsymbol{x}^*, \boldsymbol{x}_t) \leq R^2$. Now using $\eta_t = \frac{R}{\sigma\sqrt{t}}$ and $\mu_t = \frac{1}{L + \frac{1}{\eta_t}}$ we arrive at

$$\begin{aligned}\sum_{t=1}^{T}\left[\left(\gamma + \frac{1}{\gamma} - 1\right) \mathbb{E}\, F(\boldsymbol{x}_t) + \mathbb{E}\, F(\boldsymbol{x}_{t+1}) - \gamma F(\boldsymbol{x}^*)\right] &\geq -R^2\left(L + \frac{\sigma}{R}\sqrt{T}\right) - \frac{\sigma R}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \\ &\overset{(a)}{\geq} -\left(R^2 L + 2\sigma R\sqrt{T}\right).\end{aligned}$$

Here, (a) follows from the fact that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$. Thus

$$\begin{aligned}\sum_{t=1}^{T}\left(\gamma + \frac{1}{\gamma}\right)\mathbb{E}[F(\boldsymbol{x}_t)] - \gamma\mathrm{OPT} &= \sum_{t=1}^{T}\left[\left(\gamma + \frac{1}{\gamma} - 1\right) \mathbb{E}\, F(\boldsymbol{x}_t) + \mathbb{E}\, F(\boldsymbol{x}_{t+1}) - \gamma F(\boldsymbol{x}^*)\right] + F(\boldsymbol{x}_1) - F(\boldsymbol{x}_{T+1}) \\ &\geq -\left(R^2 L + 2\sigma R\sqrt{T} + \mathrm{OPT}\right).\end{aligned}$$

Dividing both sides by $\left(\gamma + \frac{1}{\gamma}\right)T$ concludes the proof.

Acknowledgements

This work was done while the authors were visiting the Simon's Institute for the Theory of Computing. The authors would like to thank Jeff Bilmes, Volkan Cevher, Maryam Fazel, Mohammad-Reza Karimi, Andreas Krause, Mario Lucic, and Andrea Montanari for helpful discussions.

References

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

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

[3] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Van Briesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

[4] R. M. Gomez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of KDD*, 2010.

[5] C. Guestrin, A. Krause, and A. P. Singh. Near-optimal sensor placements in gaussian processes. In *ICML*, 2005.

[6] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *KDD*, 2009.

[7] B. Mirzasoleiman, A. Badanidiyuru, and A. Karbasi. Fast constrained submodular maximization: Personalized data summarization. In *ICML*, 2016.

[8] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *Proceedings of Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[9] B. Mirzasoleiman, A. Karbasi, R. Sarkar, and A. Krause. Distributed submodular maximization: Identifying representative elements in massive data. In *NIPS*, 2013.

[10] A. Singla, I. Bogunovic, G. Bartok, A. Karbasi, and A. Krause. Near-optimally teaching the crowd to classify. In *ICML*, 2014.

[11] B. Kim, O. Koyejo, and R. Khanna. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*, 2016.

[12] J. Djolonga and A. Krause. From map to marginals: Variational inference in bayesian submodular models. In *NIPS*, 2014.

[13] R. Iyer and J. Bilmes. Submodular point processes with applications to machine learning. In *Artificial Intelligence and Statistics*, 2015.

[14] F. Bach. Submodular functions: from discrete to continous domains. *arXiv preprint arXiv:1511.00394*, 2015.

[15] G. Calinescu, C. Chekuri, M. Pal, and J. Vondrak. Maximizing a submodular set function subject to a matroid constraint. *SIAM Journal on Computing*, 2011.

[16] A. Bian, B. Mirzasoleiman, J. M. Buhmann, and A. Krause. Guaranteed non-convex optimization: Submodular maximization over continuous domains. *arXiv preprint arXiv:1606.05615*, 2016.

[17] M. Karimi, M. Lucic, H. Hassani, and A. Krasue. stochastic submodular maximization: The case for coverage functions. 2017.

[18] S. A. Stan, M. Zadimoghaddam, A. Krasue, and A. Karbasi. Probabilistic submodular maximization in sub-linear time. *ICML*, 2017.

[19] S. Fujishige. *Submodular functions and optimization*, volume 58. Annals of Discrete Mathematics, North Holland, Amsterdam, 2nd edition, 2005.

[20] T. Soma and Y. Yoshida. A generalization of submodular cover via the diminishing return property on the integer lattice. In *NIPS*, 2015.

[21] C. Chekuri, T. S. Jayram, and J. Vondrak. On multiplicative weight updates for concave and submodular function maximization. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 201–210. ACM, 2015.

[22] R. Eghbali and M. Fazel. Designing smoothing functions for improved worst-case competitive ratio in online optimization. In *Advances in Neural Information Processing Systems*, pages 3287–3295, 2016.

[23] J. Edmonds. Matroids and the greedy algorithm. *Mathematical programming*, 1(1):127–136, 1971.

[24] László Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. Springer, 1983.

[25] D. Chakrabarty, Y. T. Lee, Sidford A., and S. C. W. Wong. Subquadratic submodular function minimization. In *STOC*, 2017.

[26] C. Chekuri, J. Vondrák, and R.s Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, 2011.

[27] T. Soma, N. Kakimura, K. Inaba, and K. Kawarabayashi. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *ICML*, 2014.

[28] C. Gottschalk and B. Peis. Submodular function maximization on the bounded integer lattice. In *International Workshop on Approximation and Online Algorithms*, 2015.

[29] A. Ene and H. L. Nguyen. A reduction for optimizing lattice submodular functions with diminishing returns. *arXiv preprint arXiv:1606.08362*, 2016.

[30] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[31] J. Vondrak, C. Chekuri, and R. Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 783–792. ACM, 2011.

[32] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[33] P. Brucker. An $\mathcal{O}(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3):163–166, 1984.

[34] P. M. Pardalos and N. Kovoor. An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds. *Mathematical Programming*, 46(1):321–328, 1990.

[35] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

[36] M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*, 2017.

[37] M. Soltanolkotabi. Learning ReLUs via gradient descent. *arXiv preprint arXiv:1705.04591*, 2017.

A A DR-Submodular Function that Attains $\mathrm{OPT}/2+\epsilon$ on a local maximum

We first define a (coverage) submodular set function $f: 2^V \to \mathbb{R}_+$ and then show that the multilinear extension of f has the desired property. Let $V = \{1, 2, \ldots, 2k+1\}$. We consider the following subsets of V: For $i \in \{1, 2, \ldots, k\}$ let $S_i = \{i, 2k+1\}$, for $i \in \{k+1, \ldots, 2k\}$ define $S_i = \{i\}$, and finally let $S_{2k+1} = \{1, \ldots, k\} \cup \{2k+1\}$. The submodular set function $f: 2^V \to \mathbb{R}_+$ is then defined as $f(A) = |\cup_{i\in A} S_i|$ for $A \subseteq V$. It is not hard to show that f is monotone and submodular (f is a coverage function). Let $F: [0,1]^{2k+1} \to \mathbb{R}_+$ be the multilinear extension of f. We can write

$$\begin{aligned} F(\boldsymbol{x}) &= \sum_{A\subseteq V} f(A) \prod_{i\in A} x_i \prod_{i\notin A}(1-x_i) \\ &= k+1-(1-x_{2k+1})\prod_{i=1}^{k}(1-x_i) - (1-x_{2k+1})(k-\sum_{i=1}^{k} x_i) + \sum_{i=k+1}^{2k} x_i \end{aligned}$$

Now, define $\mathcal{K} = \{\boldsymbol{x} \in [0,1]^{2k+1} : \sum_{i=1}^{2k+1} x_i = k\}$. We claim that $\boldsymbol{x}_{\mathrm{loc}} = (\overbrace{1, 1, \ldots, 1}^{k}, 0, 0, \ldots, 0)$ is a local maximum. To see this, we have $\nabla F(\boldsymbol{x}_{\mathrm{loc}}) = (\overbrace{1, 1 \cdots, 1}^{2k}, 0)$. As a result, for any $y \in \mathcal{K}$:

$$\langle \nabla F(\boldsymbol{x}_{\mathrm{loc}}), y - \boldsymbol{x}_{\mathrm{loc}} \rangle = \sum_{i=1}^{2k} y_i - k \leq 0.$$

As a result, $\boldsymbol{x}_{\text{loc}}$ is a stationary point. It remains to show that in a sufficiently small neighborhood of $\boldsymbol{x}_{\text{loc}}$ inside $\mathcal{K}$, $\boldsymbol{x}_{\text{loc}}$ becomes the maximizer of F. Note that $F(\boldsymbol{x}_{\text{loc}}) = k+1$. Consider a point

$$y = (1-\epsilon_1, \cdots, 1-\epsilon_k, \epsilon_{k+1}, \epsilon_{k+2}, \cdots, \epsilon_{2k+1}), \text{ where } \epsilon_i \in [0,\epsilon] \text{ and } \sum_{i=1}^{k} \epsilon_i = \sum_{j=k}^{2k+1} \epsilon_j. \tag{A.1}$$

It is easy to see that $y \in \mathcal{K}$. We have

$$\begin{aligned}
F(y) - F(\boldsymbol{x}_{\text{loc}}) &= \sum_{j=k+1}^{2k} \epsilon_j - (1-\epsilon_{2k+1}) \prod_{i=1}^{k} \epsilon_i - (1-\epsilon_{2k+1})(k - \sum_{i=1}^{k}(1-\epsilon_i)) \\
&= \sum_{i=1}^{k} \epsilon_i - \epsilon_{2k+1} - (1-\epsilon_{2k+1})(\prod_{i=1}^{k} \epsilon_i + \sum_{i=1}^{k} \epsilon_i) \\
&\leq \epsilon_{2k+1}(\sum_{i=1}^{k} \epsilon_i - 1) \\
&\leq \epsilon_{2k+1}(k\epsilon - 1)
\end{aligned}$$

Thus, by choosing $\epsilon \leq 1/k$, we conclude that any y with form as in (A.1) has a lower function value than $\boldsymbol{x}_{\text{loc}}$. This proves that $\boldsymbol{x}_{\text{loc}}$ is a local maximum. Now, consider the vector $\boldsymbol{x}^* = (0,0,\cdots,0,\overbrace{1,1,\cdots,1}^{k+1})$. We have $F(\boldsymbol{x}_{\text{loc}})/F(\boldsymbol{x}^*) = 1/2 + 1/(2k)$. As a result, by considering a large enough k, the value of F at the local maximum $\boldsymbol{x}_{\text{loc}}$ becomes $\text{OPT}/2 + \epsilon$.

B An Example for Deficiency of the Frank-Wolfe Type Algorithm of [16] in the Stochastic Setting

Assume we want to maximize a DR-Submodular function F over a convex set $\mathcal{K}$. Assume further that $0 \in \mathcal{K}$. The Frank-Wolfe Type algorithm discussed in [16] can be briefly stated as follows (note that for simplicity we let $\alpha = 1$ and $\delta = 0$, see Algorithm 1 in [16]): Fix a (large) number T as the total number of iterations, let $\boldsymbol{x}_0 = 0$ and for $t < T$ do:

$$\boldsymbol{x}_{t+1} = x_t + \frac{1}{T} \arg\max_{v \in \mathcal{K}} \langle v, \nabla F(\boldsymbol{x}_t) \rangle. \tag{B.1}$$

When we have access to the gradients $\nabla F(\boldsymbol{x}_t)$, it is shown in [16] that this algorithm achieves $F(\boldsymbol{x}_T) \geq (1-1/e)\text{OPT}$ for large T. Assume now that we have only access to $\boldsymbol{g}_t$ which is an unbiased estimator of $\nabla F(\boldsymbol{x}_t)$. A simple stochastic version of the above algorithm would be to replace the gradient term $\nabla F(\boldsymbol{x}_t)$ in (B.1) with $\boldsymbol{g}_t$. Here, we provide a simple example to show that this stochastic version can perform arbitrary poorly. Fix an integer n and consider $n-1$ functions $F_i : [0,1]^n \to \mathbb{R}_+$, $i \in \{1, \cdots, n-1\}$, defined as $F_i(x) = \sum_{j=1}^{n} m_{i,j} x_j$. We further let $m_{i,i} = 1$, $m_{i,n} = 1/2$, and the rest of $m_{i,j}$'s are 0. Finally we let $F(x) = \mathbb{E}_{i \sim U}[F_i(x)]$, where U is assumed to be the uniform distribution over $\{1, \cdots, n-1\}$. We want to maximize F over the polytope $\mathcal{K} = \{x : \sum_{i=1}^{n} x_i \leq 1; x_i \geq 0\}$.

Assume now that instead of $\nabla F(x)$ we have access to an unbiased estimator $\nabla F_i(x)$ where $i \sim U$. Note that $\nabla F_i(x) = (m_{i,1}, m_{i,2}, \cdots, m_{i,n})$. As a result, for $i \in \{1, \cdots, n-1\}$we obtain

$\arg\max_{v\in\mathcal{K}}\langle\nabla F_i(x), v\rangle = e_i$, where e_i is the vector that has 1 at position i and 0 elsewhere. Interestingly for this example, the stochastic Frank-Wolf algorithm never makes any progress on the $n-th$ coordinate and $\boldsymbol{x}_t$ will always take 0 on the n-th coordinate. As a result, it is easy to see that for large T the algorithm will end up at $\boldsymbol{x}_\infty = (1/(n-1), 1/(n-1), \cdots, 1/(n-1), 0)$. However, we have $\boldsymbol{x}^* = (0, 0, \cdots, 0, 1)$ and $F(\boldsymbol{x}_\infty)/F(\boldsymbol{x}^*) = 2/(n-1)$ which can become arbitrarily small with n.

Let us briefly explain why conditional gradient methods do not easily admit stochastic variants. The main bottleneck is in the update step of the continuous greedy algorithm (FW). As stated above, in each iteration, FW finds a point in the constraint set $\mathcal{K}$ which has the highest inner product with the gradient and then uses this vector in order to update the current position. However, this step is not very robust to the noise. More precisely, if instead of the gradient of F we plug into the arg max a noisy (and unbiased) version of the gradient, the outcome may be far from $\boldsymbol{v}_t$. In other words, expectation and arg max are not interchangeable. It is easy to see that the above example extends to FW with any fixed natch size (i.e. when gradient is approximated by averaging a fixed number of i.i.d. samples).

C DR-submodular Functions with Large Smoothness Parameter in ℓ_2 But Reasonable Smoothness Parameter in ℓ_1

Consider a submodular set function $f: 2^V \to \mathbb{R}$ and denote its mulilinear extension by $F: [0,1]^n \to \mathbb{R}_+$ (we also let $n \triangleq |V|$). Let us first investigate how smooth is F under the ℓ_2 norm. At $\boldsymbol{x} = \mathbf{0}$ we have $[\nabla^2 F(\mathbf{0})]_{i,j} = f(\{i,j\}) - f(\{i\}) - f(\{j\})$. Thus, for $\boldsymbol{y} = (1, 1, \cdots, 1)/\sqrt{n}$ with $\|\boldsymbol{y}\|_{\ell_2} = 1$ we have $|y^T \nabla^2 F(\mathbf{0}) y| = 1/n \sum_{i,j}(f(\{i\}) + f(\{j\}) - f(\{i,j\}))$. One can easily construct a function f such that this sum takes value $O(n)$ (also many functions in practice have this property). As a result, F can be $O(n)$-smooth. However, F may become reasonably smooth in ℓ_1 norm (with smoothness parameter that is independent of the dimension).

Lemma C.1 *For a monotone submodular function f, let m_f its denote maximum singleton value of f by m, i.e., $m_f \triangleq \max_{j\in V} f(\{j\})$. Then, the multilinear extension F is m_f-smooth under the ℓ_1 norm.*

Before proving the lemma, let us remark that in many practical applications, the value of m_f is not so large (see for example the movie recommendation setting of Section 5 where m_f is less than the maximum possible rating).

Proof At any point $\boldsymbol{x} \in [0,1]^n$, the Hessian of F, denoted by $\nabla^2 F(x)$, has the following property (see [15]):

$$
\begin{aligned}
[\nabla^2 F(\boldsymbol{x})]_{i,j} &= \frac{\partial^2 F(\boldsymbol{x})}{\partial x_i \partial x_j} \\
&= F(\boldsymbol{x}; x_i, x_j \leftarrow 1) + F(\boldsymbol{x}; x_i, x_j \leftarrow 0) - F(\boldsymbol{x}; x_i \leftarrow 1, , x_j \leftarrow 0) - F(\boldsymbol{x}; x_i \leftarrow 0, , x_j \leftarrow 1) \\
&\overset{(a)}{\geq} -\max\{f(\{i\}), f(\{j\})\} \geq -m_f,
\end{aligned}
$$

where for example by $(\boldsymbol{x}; x_i, x_j \leftarrow 1)$ we mean a vector which has value 1 on its i-th and j-th coordinate and is equal to $\boldsymbol{x}$ elsewhere. Also, (a) is a direct consequence of the submodularity of f. As a result, each element of the Hessian is negative but greater $-m$ and for any vector $\boldsymbol{y} \in \mathbb{R}^n$ we have $|\boldsymbol{y}^T \nabla^2 F(\boldsymbol{x}) \boldsymbol{y}| \leq m_f \|\boldsymbol{y}\|_{\ell_1}^2$. Hence, F is m_f-smooth under the ℓ_1 norm. ∎

D How to Construct an Unbiased Estimator of the Gradient in Multilinear Extensions

Recall that $F(\boldsymbol{x}) = \mathbb{E}_{\theta\sim\mathcal{D}}[F_\theta(\boldsymbol{x})]$. So $\nabla F_\theta(\boldsymbol{x})$ is an unbiased estimator of $\nabla F(\boldsymbol{x})$ when $\theta \sim \mathcal{D}$. Note that F_θ is a multilinear extension. It remains to provide an unbiased estimator for a generic multilinear extension $G(\boldsymbol{x})$. We have $G(\boldsymbol{x}) = \sum_{S\subseteq V}\prod_{i\in S} x_i \prod_{j\notin S}(1-x_j)g(S)$ where g is a set function. Now, it can easily be shown that

$$\frac{\partial G}{\partial x_i} = G(\boldsymbol{x}; x_i \leftarrow 1) - G(\boldsymbol{x}; x_i \leftarrow 0).$$

where for example by $(\boldsymbol{x}; x_i \leftarrow 1)$ we mean a vector which has value 1 on its i-th coordinate and is equal to $\boldsymbol{x}$ elsewhere. To create an unbiased estimator for $\frac{\partial G}{\partial x_i}$ at a point $\boldsymbol{x}$ we can simply sample a set S by including each element in it independently with probability x_i and use $g(S\cup\{i\})-g(S\setminus\{i\})$ as an unbiased estimator for the i-th partial derivative. We can sample one single set S and use the above trick for all the coordinates. This involves n function computations for g. Having a batch size B we can repeat this procedure B times and then average.