Supplementary material for "Online Submodular Minimization for Combinatorial Structures"

Here, we provide details of proofs in the main paper.

1. Proof of Lemma 1

Lemma 1. *Let g_t be a subgradient of f_t (obtained by the greedy algorithm). Then $\|g_t\| \leq \beta \max_{A\subseteq E} |f_t(A)| - f_t(\emptyset)$, where $\beta = 1$ if f_t is nondecreasing, and $\beta = 3$ otherwise.*

Proof. Since t is fixed, we will drop the subscript in this proof. Essential for the proof is that $g \in P_f$, in fact, it lies in the base polytope (Fujishige, 2005). This means that

$$g \cdot \chi_A \leq f(A) \tag{1}$$

for all $A \subseteq E$. Assume first that f is nonnegative and nondecreasing. Then Equation (1) immediately leads to a bound on $\|g\|$, by bounding the ℓ_2 norm by the ℓ_1 norm:

$$\|g\|_2 \leq \|g\|_1 = g \cdot \chi_E \leq f(E). \tag{2}$$

This proves the lemma for nondecreasing functions.

For arbitrary submodular functions, we use the construction of g in slightly more detail, but the basic arguments are the same. For ease of notation, let $\gamma = \max_{A\subseteq E} |f(A)|$. We first recall how g was constructed, given $x \geq 0$. We denote the components of x by x_i, $1 \leq i \leq m$. We find a permutation π such that $x_{\pi(1)} \geq x_{\pi(2)} \geq \ldots \geq x_{\pi(m)}$. This ordering induces a maximal chain of sets, $\emptyset = A_0 \subset A_1 \subset \ldots \subset A_m$ with $A_0 = \emptyset$ and $A_i = A_{i-1} \cup \{e_{\pi(i)}\}$. Setting

$$g_{\pi(i)} = f(A_i) - f(A_{i-1}) \tag{3}$$

yields g, with $g \cdot \chi_{A_i} = f(A_i) - f(\emptyset)$. Let $g^+ = \max\{g, 0\}$ be the element-wise maximum.

Claim 1. $\|g^+\|_1 = \sum_{i=1}^M g_i^+ \leq \gamma - f(\emptyset)$.

Consider the subset E^+ of elements e_k with $g_k \geq 0$, and let B_j be the set of the j first such elements, where we use the ordering induced by x, restricted to E^+. We call this restriction π^+: $\{1, \ldots |E^+|\} \to E^+$. The jth element in E^+, $e_{\pi^+(j)}$, also occurs at some point $\pi(i(j))$ in the full sequence, so that $e_{\pi^+(j)} = e_{\pi(i(j))}$. Since the nonnegative elements are a subsequence, we know that $i(j) \geq j$ and thus $B_j \subseteq A_{i(j)}$. By the definition of g_j and diminishing marginal costs (submodularity), it holds for all $e_{\pi'(j)} \in E^+$ that

$$\begin{aligned} g_{\pi'(j)} &= f(A_{i(j)-1} \cup \{e_j\}) - f(A_{i(j)-1}) && (4)\\ &\leq f(B_{j-1} \cup \{e_j\}) - f(B_j) =: g'_j. && (5)\end{aligned}$$

The definition of the g'_j implies that $\sum_j g'_j + f(\emptyset) = f(B_j) \leq \gamma$. In consequence,

$$\|g^+\|_1 = \sum_{j=1}^{|E^+|} g_{\pi'(j)} \leq \sum_{j=1}^{|E^+|} g'_j \leq \gamma - f(\emptyset). \tag{6}$$

This proves the claim.

We next use this result to bound the sum of the absolute values of the negative entries. By definition of g, we know that $\sum_{i=1}^k g_i = f(A_k) - f(\emptyset)$, in particular also for $k = m$. Since $-\gamma \leq f(A_k) \leq \gamma$, it follows that

$$-\gamma - f(\emptyset) \leq \sum_{i:g_i<0} g_i + \sum_{i:g_i\geq 0} g_i \leq \gamma - f(\emptyset). \tag{7}$$

Using the claim, we get that

$$\sum_{i:g_i<0} g_i \geq -\gamma - f(\emptyset) - \sum_{i:g_i\geq 0} g_i \geq -2\gamma, \tag{8}$$

and thus $\sum_{i:g_i<0} |g_i| \leq 2\gamma$. In total, this shows that

$$\|g\|_1 = \sum_{i:g_i<0} |g_i| + \sum_{i:g_i\geq 0} |g_i| \leq 3\gamma - f(\emptyset) \leq 4\gamma. \tag{9}$$

With $\|g\|_2 \leq \|g\|_1$, the lemma follows. □

2. Detailed proof of Theorem 2

First, we re-state the theorem.

Theorem 2. *For an approximation $\hat{f}$ that satisfies (C1) and (C2), $M = \max_t f_t(E)$, and $\eta = T^{-1/2}$, Algorithm 2 achieves an expected α-regret $\mathbb{E}[R_\alpha(T)] \leq 3\alpha m M/\sqrt{T} = O(\alpha m/\sqrt{T})$.*

Proof. Let

$$S_t = \operatorname*{argmin}_{S\in\mathcal{S}} \sum\nolimits_{\tau=1}^{t-1} \hat{f}_\tau(S) + \alpha r(S);$$

$$\widehat{S}_t = \operatorname*{argmin}_{S\in\mathcal{S}} \sum\nolimits_{\tau=1}^{t-1} \hat{f}_\tau(S); \quad S_t^* = \operatorname*{argmin}_{S\in\mathcal{S}} \sum\nolimits_{\tau=1}^{t} f_\tau(S).$$

First, we show a relation for $\sum_{t=1}^{T} \hat{f}_t(S_{t+1})$ and later relate it to the actual cost $\sum_{t=1}^{T} \hat{f}_t(S_t)$. The first inequality is

$$\sum_{t=1}^{T} \hat{f}_t(\widehat{S}_{t+1}) \leq \sum_{t=1}^{T} \hat{f}_t(\widehat{S}_{T+1}). \tag{10}$$

It holds trivially for $T = 1$. The case $T + 1$ follows by induction and the optimality of $\widehat{S}_{T+1}$:

$$\begin{aligned}\sum_{t=1}^{T+1} \hat{f}_t(\widehat{S}_{t+1}) &\leq \sum_{t=1}^{T} \hat{f}_t(\widehat{S}_{T+1}) + \hat{f}_{T+1}(\widehat{S}_{T+2}) \\ &\leq \sum_{t=1}^{T} \hat{f}_t(\widehat{S}_{T+2}) + \hat{f}_{T+1}(\widehat{S}_{T+2}) \\ &= \sum_{t=1}^{T+1} \hat{f}_t(\widehat{S}_{T+2}).\end{aligned}$$

We now replace $\hat{f}_1$ in Equation (10) by $\hat{f}_1 + \alpha r$:

$$\begin{aligned}\sum_{t=1}^{T} \hat{f}_t(S_{t+1}) + \alpha r(S_1) &\leq \sum_{t=1}^{T} \hat{f}_t(S_{T+1}) + \alpha r(S_{T+1}) \\ &\leq \sum_{t=1}^{T} \hat{f}_t(\widehat{S}_{T+1}) + \alpha r(\widehat{S}_{T+1}).\end{aligned}$$

Rearranging the terms yields

$$\sum_{t=1}^{T} \hat{f}_t(S_{t+1}) \leq \sum_{t=1}^{T} \hat{f}_t(\widehat{S}_{T+1}) + \alpha(r(\widehat{S}_{T+1}) - r(S_1)). \tag{11}$$

To transfer this result to the series of S_t, we use that $\hat{f}_t(S_t) \leq \hat{f}_t(S_{t+1}) + (\hat{f}_t(S_t) - \hat{f}_t(S_{t+1}))$:

$$\begin{aligned}\sum_{t=1}^{T} \hat{f}_t(S_t) \leq \sum_{t=1}^{T} \hat{f}_t(\widehat{S}_{T+1}) \\ + \sum_{t=1}^{T} (\hat{f}_t(S_t) - \hat{f}_t(S_{t+1})) + \alpha(r(\widehat{S}_{T+1}) - r(S_1)).\end{aligned} \tag{12}$$

Condition (C1) implies that

$$\sum_{t=1}^{T} \hat{f}_t(\widehat{S}_{T+1}) \leq \sum_{t=1}^{T} \hat{f}_t(S_T^*) \leq \alpha \sum_{t=1}^{T} f_t(S_T^*),$$

and that $\sum_{t=1}^{T} f_t(S_t) \leq \sum_{t=1}^{T} \hat{f}_t(S_t)$. Together with Equation (12), this yields

$$\begin{aligned}\sum_{t=1}^{T} f_t(S_t) - \alpha \sum_{t=1}^{T} f_t(S_T^*) \\ \leq \sum_{t=1}^{T} (\hat{f}_t(S_t) - \hat{f}_t(S_{t+1})) + \alpha(r(\widehat{S}_{T+1}) - r(S_1)).\end{aligned} \tag{13}$$

It remains to bound the two terms on the right hand side, and these bounds depend on $r \in [0, M/\eta]^E$.

We first address the random perturbation r in $[0, M/\eta]^E$. The last term can be bounded as

$$\alpha \mathbb{E}[r(\widehat{S}_{T+1}) - r(S_1)] \leq \alpha m M/\eta. \tag{14}$$

To bound the expected sum of differences of the function values, we use a technique by Hazan & Kale (2009). For the analysis, one can assume that r is resampled in each round. We first bound $P(S_t \neq S_{t+1})$. A simple union bound holds:

$$\begin{aligned}P(S_t \neq S_{t+1}) \leq \sum_{i=1}^{m} P(e_i \in S_t \text{ and } e_i \notin S_{t+1}) \\ + \sum_{i=1}^{m} P(e_i \notin S_t \text{ and } e_i \in S_{t+1}).\end{aligned} \tag{15}$$

To bound the right hand side, we fix i and look at $P(e_i \in S_t \text{ and } e_i \notin S_{t+1})$. Denote the components of r by r_j and define $r' : 2^E \to \mathbb{R}$ as $r'(S) = \sum_{e_j \in S, j \neq i} r_j$, so $r'(e_j) = r(e_j) = r_j$ for all $j \neq i$, but $r'(e_i) = 0$; and define $\Phi_t' : 2^E \to \mathbb{R}$ as $\Phi_t'(S) = \sum_{\tau=1}^{t-1} \hat{f}_\tau + \alpha r'(S)$. Now let

$$S^1 = \operatorname*{argmin}_{S \in \mathcal{S}, e_i \in S} \Phi_t'(S); \qquad S^2 = \operatorname*{argmin}_{S \in \mathcal{S}, e_i \notin S} \Phi_t'(S).$$

The event $e_i \in S_t$ only happens if $\Phi_t'(S^1) + \alpha r_i < \Phi_t'(S^2)$ and $S_t = S^1$. On the other hand, to have $e_i \notin S_{t+1}$, it must be that $\Phi_t'(S^1) + \alpha r_i \geq \Phi_t'(S^2) - \alpha M$, since otherwise

$$\begin{aligned}\sum_{\tau=1}^{t+1} \hat{f}_\tau(S^1) + \alpha r(S^1) &= \Phi_t'(S^1) + \alpha r_i + \hat{f}_t(S^1) \\ &< \Phi_t'(S^2) \\ &< \Phi_t'(B) + \hat{f}_t(B)\end{aligned}$$

for all $B \in \mathcal{S}$ with $e_i \notin B$. Here, we used that $\hat{f}_t(S) \leq \alpha f_t(S) \leq \alpha M$ for all $S \subseteq E$. Let $v = \alpha^{-1}(\Phi'(S^2) - \Phi'(S^1))$, then $e_i \in S_t$ and $e_i \notin S_{t+1}$ only if $r_i \in [v - M, v]$. The number r_i is in this range with probability at most η since it is chosen uniformly at random from $[0, M/\eta]$, so $P(e_i \in S_t \text{ and } e_i \notin S_{t+1}) \leq \eta$. The bound on $P(e_i \notin S_t \text{ and } e_i \in S_{t+1})$ follows by an analogous argumentation. Together, those results bound (15):

$$\begin{aligned}P(S_t \neq S_{t+1}) &\leq \sum_{i=1}^{m} P(e_i \in S_t \text{ and } e_i \notin S_{t+1}) \\ &\quad + \sum_{i=1}^{m} P(e_i \notin S_t \text{ and } e_i \in S_{t+1}) \\ &\leq 2m\eta.\end{aligned} \tag{16}$$

Equation 16 helps to bound the sum of function values, using $\hat{f}(C) \leq \alpha M$ for all C:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\big[\hat{f}_t(S_t) - \hat{f}_t(S_{t+1})\big] \\ \leq \sum_{t=1}^{T} P(S_t \neq S_{t+1}) \max_{B \in \mathcal{S}} \hat{f}(B) \\ \leq 2\alpha m M T \eta.\end{aligned} \tag{17}$$

Combining Inequalities (13), (14) and (17) results in

$$\begin{aligned}\mathbb{E}\big[\sum_{t=1}^{T} f_t(S_t)\big] - \alpha \sum_{t=1}^{T} f_t(S_T^*) \\ \leq \alpha M m/\eta + 2\alpha m M T \eta.\end{aligned}$$

The final regret bound follows for $\eta = T^{-1/2}$. □

3. Proof of Lemma 2

Lemma 2. *Let $\hat{f}$ be randomly chosen between $\hat{f}^-$ and $\hat{f}^+$ with equal probabilities. Then $f(S) \leq \mathbb{E}[\hat{f}(S)] \leq (|V|/2)f(S)$ for all minimal (s,t)-cuts S.*

Proof. First, we bound $\hat{f}^-(S)$. Let $\Delta^-(S)$ be the set of head nodes of edges in S, i.e., at most all nodes on the t side of the cut.

$$\begin{aligned}\hat{f}^-(S) &= \sum\nolimits_{v\in\Delta^-(S)} f(S\cap E_v^-)\\ &\leq |\Delta^-(S)| \max_{v\in\Delta^-(S)} f(S\cap E_v^-)\\ &\leq |\Delta^-(S)| f(S).\end{aligned}$$

Analogously, it follows that $\hat{f}^+(S) \leq |\Delta^+(S)|f(S)$, $|\Delta^+(S)|$ being the number of tail nodes of edges in S. We combine these bounds to

$$\begin{aligned}\mathbb{E}[\hat{f}(S)] &= (\hat{f}^-(S) + \hat{f}^+(S))/2\\ &\leq f(S)(|\Delta^+(S)| + |\Delta^-(S)|)/2\\ &\leq f(S)|V|/2.\end{aligned}$$

□

4. Proof of Lemma 3

Let $S^* = \operatorname{argmin}_{S\in\mathcal{S}} \sum_t f_t(S)$, and $\hat{S}^* = \operatorname{argmin}_{S\in\mathcal{S}} \sum_t \hat{f}_t^2(S)$. We play S_t as prescribed by algorithm $\mathcal{A}$.

Lemma 3. *Let $\widehat{R}_{\mathcal{A}}$ be the regret of an online algorithm $\mathcal{A}$ when used with linear cost functions with a range like $\hat{f}_t^2$. Using $\mathcal{A}$ with $\hat{f}_t^2$ when observing f_t leads to an α_g-regret of $R_{\alpha_g}(T) \leq \alpha_g \widehat{R}_{\mathcal{A}}/\nu$.*

Proof. Since we use $\hat{f}_t^2$ in $\mathcal{A}$, the regret $\widehat{R}_{\mathcal{A}}$ bounds $\sum_t(\hat{f}_t^2(S_t) - \hat{f}_t^2(\widehat{S}^*))$. Therefore, we relate the actual regret, $\sum_t(f_t(S_t) - \alpha_g f_t(S^*))$, to the regret of $\mathcal{A}$. We use that $\hat{f}^2(S) \leq f^2(S) \leq \alpha_g^2 \hat{f}^2(S)$. We have that

$$\begin{aligned}\sum\nolimits_t (f_t(S_t) - \alpha_g f_t(S^*)) &= \sum\nolimits_t \frac{(f_t^2(S_t) - \alpha_g^2 f_t^2(S^*))}{(f_t(S_t) + \alpha_g f_t(S^*))}\\ &\leq \sum\nolimits_t (f_t^2(S_t) - \alpha_g^2 f_t^2(S^*))/(\alpha_g\nu)\\ &\leq \sum\nolimits_t \alpha_g^2(\hat{f}_t^2(S_t) - \hat{f}_t^2(S^*))/(\alpha_g\nu)\\ &\leq \sum\nolimits_t \alpha_g(\hat{f}_t^2(S_t) - \hat{f}_t^2(\widehat{S}^*))/(\nu)\\ &= \alpha_g \sum\nolimits_t \widehat{R}_{\mathcal{A}}/\nu,\end{aligned}$$

since $\hat{S}^* = \operatorname{argmin}_{S\in\mathcal{S}} \sum_t \hat{f}_t^2(S)$ is optimal for $\hat{f}^2$. □

5. Multiple labels for label costs in Algorithm 3

Here, we outline how to simulate label costs when one edge can have more than one label. This simulation applies to the spanning tree example.

Let k be the maximum number of labels any edge can have. We assign k "slots" to each edge. Each label $\ell \in \pi(e)$ occupies $1 \leq \gamma_e(\ell) \leq k$ slots, such that $\sum_{\ell\in\pi(e)} \gamma_e(\ell) = k$. Define k copies $G_i = (V, E_i)$ of G. Edge e is contained in $E_i(L)$ if i of its slots are filled by labels in L. Then we use

$$g(L) = \sum\nolimits_{i=1}^{k} r(E_i(L)).$$

This sum is still submodular, and maximum only if $E(L)$ contains a tree of full edges. The approximation factor increases moderately to $O(\log(nk))$.

6. Problems when applying the algorithms in (Kakade et al., 2009) to the submodular-cost setting

Kakade et al. (2009) show online approximation algorithms that use an offline approximation algorithm as a black box. Their method generalizes online gradient descent (Zinkevich, 2003) to use the approximation algorithm in an approximate projection. Their cost function is of the form $c : 2^E \times \mathbb{R}^d \to \mathbb{R}$, $c(S, w) = \langle\phi(S), w\rangle$ and must be *linear* in w. That means, it is the dot product between some feature vector of S and a weight vector. (In the paper, they leave nonlinear costs as an open problem.)

To use this framework, we must express any non-decreasing submodular f via a cost vector w^f as $c(S, w^f) = f(S)$. The set of non-decreasing submodular functions on E is equivalent to a convex cone in $\mathbb{R}^{2^{|E|}}$. This set has a non-empty relative interior (e.g., $f(S) = \log(1 + |S|)$). As a result, simple linear algebra shows that a full basis is needed to represent all such f meaning that w has an exponential dimension d. But then the regret bound in (Kakade et al., 2009) is exponential in $|E|$, since it is linear in $\|w\|$, i.e., proportional to $\sqrt{d}$. Whilst the norm issue can possibly be resolved, the algorithm also assumes that, given any $w \in \mathbb{R}^d$, we can project it onto the set of those w for which $c(\cdot, w)$ is a nondecreasing submodular function. Given the results in (Seshadri & Vondrák, 2010), this too seems to be non-trivial.

References

Fujishige, S. *Submodular Functions and Optimization.* Number 58 in Annals of Discrete Mathematics. Elsevier Science, 2nd edition, 2005.

Hazan, E. and Kale, S. Online submodular minimization. In *Proc. of the Ann. Conf. on Neural Info. Processing Systems (NIPS)*, 2009.

Kakade, S., Kalai, A. T., and Ligett, K. Playing games with approximation algorithms. *SIAM Journal on Computing*, 39(3):1088–1106, 2009.

Seshadri, C. and Vondrák, J. Is submodularity testable? In *arXiv 1008.0831v1*, 2010.

Zinkevich, M. Online convex programming and infinitesimal gradient ascent. In *Proc. of the Int. Conf. on Machine Learning (ICML)*, 2003.