MA1407.00050 Mathematical Algorithms II **SNU, Fall 2025**
Lecture #16: SAM and Submodular Function Minimization November 4, 2025
Lecturer: Dabeen Lee

1 Outline

In this lecture, we cover

- sharpness-aware minimization (SAM),
- submodular function minimization

2 Sharpness-Aware Minimization

The next question that we consider is about generalization. It has been observed that flat minima

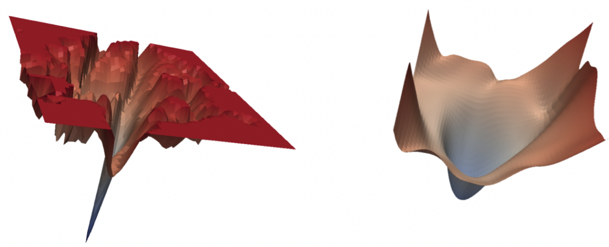

Figure 16.1: Sharp and flat minima

tend to generalize better than sharp minima. Rather than finding a single θ that has a low loss, we look for θ such that parameter θ' contained in the neighborhood of θ uniformly has a low loss value. To achieve this goal, we consider the following formulation.

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n} \max_{\delta:\|\delta\|_2\leq \epsilon} \quad \ell\left(f_{\theta+\delta}(x_i), y_i\right).$$

This framework is **sharpness-aware minimization (SAM)** [FKMN21]. Here, the inner maximization can be approximated using the first-order Taylor approximation.

$$\max_{\delta:\|\delta\|_2\leq \epsilon} \ell\left(f_{\theta+\delta}(x_i), y_i\right) \simeq \max_{\delta:\|\delta\|_2\leq \epsilon} \left\{\ell\left(f_{\theta}(x_i), y_i\right) + \delta^\top \nabla_\theta \ell\left(f_\theta(x_i), y_i\right)\right\}.$$

The maximizer of the approximation is simply given by

$$\delta^* = \frac{\epsilon}{\|\nabla_\theta \ell\left(f_\theta(x_i), y_i\right)\|_2} \nabla_\theta \ell\left(f_\theta(x_i), y_i\right).$$

Then we substitute $\delta = \delta^*$ and consider $f_{\theta+\delta^*}$:

$$\ell\left(f_{\theta+\delta^*}(x_i), y_i\right).$$

Here,

$$\nabla_\theta\left(\ell\left(f_{\theta+\delta^*}(x_i), y_i\right)\right) = \left(1 + \frac{d\delta^*}{d\theta}\right)\nabla_\theta \ell\left(f_{\theta+\delta^*}(x_i), y_i\right) \simeq \nabla_\theta \ell\left(f_{\theta+\delta^*}(x_i), y_i\right).$$

3 Submodular Functions

Let E be a set of elements. We say that a set function $f : 2^E \to \mathbb{R}$ is **submodular** if it satisfies

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \quad \text{for all } A, B \subseteq E.$$

An equivalent definition of submodularity for set functions is the notion of **diminishing marginal returns** property. That is, a set function $f : 2^E \to \mathbb{R}$ is submodular if and only if it satisfies

$$f(A \cup \{e\}) - f(A) \geq f(B \cup \{e\}) - f(B) \quad \text{for all } A \subseteq B \subseteq E \text{ and } e \notin B.$$

Many functions that arise in discrete and combinatorial optimization problems turn out to be submodular. Let us provide a few representative examples below.

- **Linear function:** For any $w \in \mathbb{R}^{|E|}$, f with $f(S) = \sum_{e \in S} w_e$ for $S \subseteq E$ is submodular.
- **Concave utility:** For any concave function $g : \mathbb{R}_+ \to \mathbb{R}$ and $w \in \mathbb{R}_+^{|E|}$, f with $f(S) = g(\sum_{e \in S} w_e)$ for $S \subseteq E$ is a submodular function.
- **Coverage function:** Suppose that each element $e \in E$ corresponds to some area A_e. Then f with $f(S) = |\cup_{e \in S} A_e|$ for $S \subseteq E$ is submodular.
- **Success probability:** Let $p_e \in [0, 1]$ for $e \in E$. Then f with $f(S) = 1 - \prod_{e \in S}(1 - p_e)$ for $S \subseteq E$ is submodular.
- **Graph cuts:** Let $G = (V, E)$ be an undirected graph. Then f with $f(S) = |\delta(S)|$ for $S \subseteq V$ is submodular, where $\delta(S)$ is the set of edges crossing the partition $(S, V \setminus S)$ of the vertex set V.
- **Directed cuts:** Let $D = (N, A)$ be a directed graph. Then f with $f(S) = |\delta^+(S)|$ for $S \subseteq V$ is submodular, where $\delta^+(S)$ is the set of arcs leaving S.
- **Matroid rank functions:** Let $\mathcal{M} = (E, \mathcal{I})$ be a matroid. Then its rank function r given by $r(S) = \max\{|A| : A \in \mathcal{I}\}$ for $S \subseteq E$ is submodular.

As this wide range of examples suggests, submodular functions provide a useful framework for modeling discrete-valued decision variables. For utility, coverage, and success probability functions, the problem of maximizing a submodular function is relevant. For cut functions, submodular function minimization is relevant. As a first step, in this lecture, we consider the minimization problem.

4 Submodular Function Minimization

Let us consider the problem of minimizing a submodular function. Given a submodular function $f : 2^E \to \mathbb{R}$ over the element set E, we consider

$$\text{minimize} \quad f(S) \quad \text{subject to} \quad S \subseteq E. \tag{16.1}$$

Since f is a set function, we can interpret the function over the set of binary vectors $\{0, 1\}^{|E|}$. To be more precise, any $S \subseteq E$ can be represented by its characteristic vector $\mathbf{1}_S \in \{0, 1\}^{|E|}$ that takes

1 for the elements in S and 0 for the other elements. Similarly, any vector $z \in \{0,1\}^{|E|}$ corresponds to a subset $S_z = \{e \in E : z_e = 1\}$. Then, with a slight abuse of notation, we may define

$$f(z) := f(S_z).$$

In this case, (16.1) can be rewritten as the following binary optimization problem:

$$\text{minimize} \quad f(z) \quad \text{subject to} \quad z \in \{0,1\}^{|E|}. \tag{16.2}$$

4.1 Lovász Extension

We will show that problem (16.2), and thus problem (16.1), can be reformulated as a convex optimization problem. To this end, we introduce the **Lovász extension** of a submodular function. Given any function $f : \{0,1\}^{|E|} \to \mathbb{R}$, its Lovász extension $\hat{f} : [0.1]^{|E|} \to \mathbb{R}$ is defined as

$$\hat{f}(x) = \sum_{i=0}^{|E|} \lambda_i f(\mathbf{1}_{S_i})$$

where $\emptyset = S_0 \subset S_1 \subset \cdots \subset S_{|E|} = E$ and $\lambda_i \geq 0$ with $\sum_{i=0}^{|E|} \lambda_i = 1$ such that

$$x = \sum_{i=0}^{|E|} \lambda_i \mathbf{1}_{S_i}.$$

Let $\pi(i) = S_i \setminus S_{i-1}$ for $i = 1, 2, \ldots, |E|$. Then we have

$$S_i = \{\pi(1), \pi(2), \ldots, \pi(i)\} \quad \text{for } i = 1, 2, \ldots, |E|,$$

and moreover, it follows from the definition that

$$\begin{aligned} x_{\pi(1)} &= \lambda_1 + \lambda_2 + \cdots + \lambda_{|E|}, \\ x_{\pi(2)} &= \lambda_2 + \cdots + \lambda_{|E|}, \\ &\vdots \\ x_{\pi(|E|)} &= \lambda_{|E|}. \end{aligned}$$

This identifies the coefficients λ_i as

$$\lambda_i = x_{\pi(i)} - x_{\pi(i+1)} \quad \text{for } i = 0, 1, 2, \ldots, |E|,$$

with the convention that $x_{\pi(0)} = 1$ and $x_{\pi(|E|+1)} = 0$. To summarize, the Lovász extension can be expressed as

$$\hat{f}(x) = \sum_{i=0}^{|E|} (x_{\pi(i)} - x_{\pi(i+1)}) f(\mathbf{1}_{S_i}).$$

An equivalent definition that avoids the explicit use of the coefficients λ_i is given as follows. For any $x \in [0,1]^{|E|}$, we sort the components of x in non-increasing order with a permutation $\pi : \{1, 2, \ldots, |E|\} \to \{1, 2, \ldots, |E|\}$ such that

$$x_{\pi(1)} \geq x_{\pi(2)} \geq \cdots \geq x_{\pi(|E|)}.$$

Then, we define $S_0 = \emptyset$ and $S_i = \{\pi(1), \pi(2), \ldots, \pi(i)\}$ for $i = 1, 2, \ldots, |E|$ and set $x_{\pi(0)} = 1$ and $x_{\pi(|E|+1)} = 0$. With these notations, the Lovász extension can be expressed as

$$\hat{f}(x) = \sum_{i=0}^{|E|} (x_{\pi(i)} - x_{\pi(i+1)}) f(\mathbf{1}_{S_i}).$$

Basically, the Lovász extension $\hat{f}(x)$ is defined as the convex combination of the function values $f(\mathbf{1}_{S_i})$ where the convex coefficients λ_i are chosen so that x is represented as the convex combination of the characteristic vectors $\mathbf{1}_{S_i}$. The following provides another equivlanet definition of the Lovász extension.

Proposition 16.1. *Let $f : \{0,1\}^{|E|} \to \mathbb{R}$. Then, for any $x \in [0,1]^{|E|}$, its Lovász extension $\hat{f}$ satisfies that*

$$\hat{f}(x) = \mathbb{E}_{\lambda \sim U[0,1]} \left[f\left(\{e \in E : x_e \geq \lambda\}\right)\right]$$

where $U[0,1]$ is the uniform distribution on the interval $[0,1]$.

Proof. Let $x \in [0,1]^{|E|}$. Take a permutation $\pi : \{1, 2, \ldots, |E|\} \to \{1, 2, \ldots, |E|\}$ such that

$$x_{\pi(1)} \geq x_{\pi(2)} \geq \cdots \geq x_{\pi(|E|)}.$$

Then, we have

$$\hat{f}(x) = \sum_{i=0}^{|E|} (x_{\pi(i)} - x_{\pi(i+1)}) f(\mathbf{1}_{S_i})$$

where $S_0 = \emptyset$ and $S_i = \{\pi(1), \pi(2), \ldots, \pi(i)\}$ for $i = 1, 2, \ldots, |E|$ and $x_{\pi(|E|+1)} = 0$. For simpler presentations, we assume that

$$1 > x_{\pi(1)} > x_{\pi(2)} > \cdots > x_{\pi(|E|)} > 0.$$

Take $\lambda \sim U[0,1]$. Then we have $\{e \in E : x_e \geq \lambda\} = S_i$ if and only if $x_{\pi(i)} \geq \lambda > x_{\pi(i+1)}$ for $i = 0, 1, \ldots, |E|$, with the convention that $x_{\pi(0)} = 1$ and $x_{\pi(|E|+1)} = 0$. Therefore, it follows that $\hat{f}(x) = \mathbb{E}_{\lambda \sim U[0,1]} \left[f\left(\{e \in E : x_e \geq \lambda\}\right)\right]$, as required. □

References

[FKMN21] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 2