

Advanced Topics in Data Sciences

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 1: Course Overview and Introduction to Submodularity

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2015)

lions@epfl



License Information for Mathematics of Data Slides

- ▶ This work is released under a [Creative Commons License](#) with the following terms:
- ▶ **Attribution**
 - ▶ The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- ▶ **Non-Commercial**
 - ▶ The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- ▶ **Share Alike**
 - ▶ The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- ▶ [Full Text of the License](#)

Outline

▶ This lecture

1. Overview of the course
2. Recap of compressive sensing
3. Subsampled measurement matrices
4. Structured sparsity
5. Learning-based subsampling
6. Submodularity: Definitions, Properties, and Examples

Prerequisites

Students are expected to have a good understanding of the basics of the following areas of mathematics:

- ▶ Probability
- ▶ Linear algebra
- ▶ Calculus

Familiarity with *convex sets* and *convex functions* will also be highly beneficial.

Further Information

The hours of the course are as follows:

- ▶ Lectures on Fridays, 14:15-16:00 in ELG116
- ▶ Office hours *by appointment only* – contact Prof. Volkan Cevher at volkan.cevher@epfl.ch

The grading of the course is based on the following:

- ▶ Part of the grade will be based on *attendance* (1 point).
- ▶ Each student will be asked to *scribe one or two lectures* (2 points).
- ▶ Each student will *complete a project and present it* (3 points).

Course Objectives

The broad goal of this course is to present theory and methods for addressing key challenges in modern data sciences.

The content of the course is split into three related areas:

- ▶ Discrete optimization (weeks 1–4)
- ▶ Convex optimization (weeks 5–8)
- ▶ Statistical learning theory (weeks 9–12)

Questions?

Part I: Learning-based Compressive Subsampling

Recommended reading:

- ▶ *An introduction to compressive sampling*, Candès and Wakin, 2008
- ▶ *Learning-based compressive subsampling*, Baldassarre, Li, Scarlett, Gözcü, Bogunovic, and Cevher, 2012

Signal recovery from linear measurements

The following problem is fundamental in signal processing, machine learning, and many other areas.

Estimation from linear measurements: Problem statement

Recover an accurate estimate $\hat{\mathbf{x}}$ of a signal $\mathbf{x} \in \mathbb{C}^p$ from a set of linear measurements of the form

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{w},$$

where $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a *known* measurement matrix, and $\mathbf{w} \in \mathbb{C}^{n \times 1}$ is additive noise.

Signal recovery from linear measurements

The following problem is fundamental in signal processing, machine learning, and many other areas.

Estimation from linear measurements: Problem statement

Recover an accurate estimate $\hat{\mathbf{x}}$ of a signal $\mathbf{x} \in \mathbb{C}^p$ from a set of linear measurements of the form

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{w},$$

where $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a *known* measurement matrix, and $\mathbf{w} \in \mathbb{C}^{n \times 1}$ is additive noise.

Examples:

- ▶ Image compression
- ▶ Medical resonance imaging (MRI)
- ▶ Communications
- ▶ Linear regression

Signal recovery from linear measurements

The following problem is fundamental in signal processing, machine learning, and many other areas.

Estimation from linear measurements: Problem statement

Recover an accurate estimate $\hat{\mathbf{x}}$ of a signal $\mathbf{x} \in \mathbb{C}^p$ from a set of linear measurements of the form

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{w},$$

where $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a *known* measurement matrix, and $\mathbf{w} \in \mathbb{C}^{n \times 1}$ is additive noise.

Examples:

- ▶ Image compression
- ▶ Medical resonance imaging (MRI)
- ▶ Communications
- ▶ Linear regression

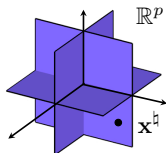
Two regimes of interest:

- ▶ $n > p$ (*overdetermined*): Solvable using classical techniques such as least squares
- ▶ $n < p$ (*underdetermined*): Infinitely many solutions; impossible in general

A natural signal model

Definition (s -sparse vector)

A vector $\alpha \in \mathbb{R}^p$ is s -sparse if it has at most s non-zero entries.

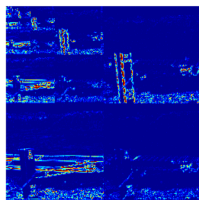


$$\mathbf{x}^h = \Phi \alpha^h$$

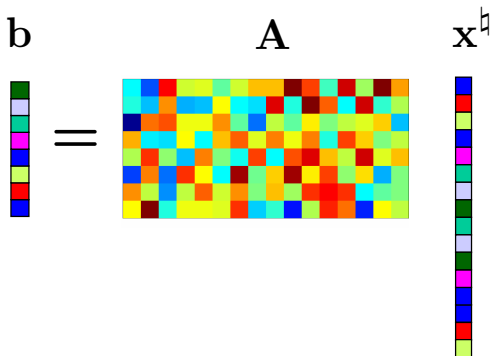
Sparse representations

α^h : *sparse* transform coefficients

- ▶ Basis representations $\Phi \in \mathbb{R}^{p \times p}$
 - ▶ *Wavelets*, DCT, ...
- ▶ Frame representations $\Phi \in \mathbb{R}^{m \times p}$, $m > p$
 - ▶ Gabor, curvelets, shearlets, ...
- ▶ Other *dictionary* representations...

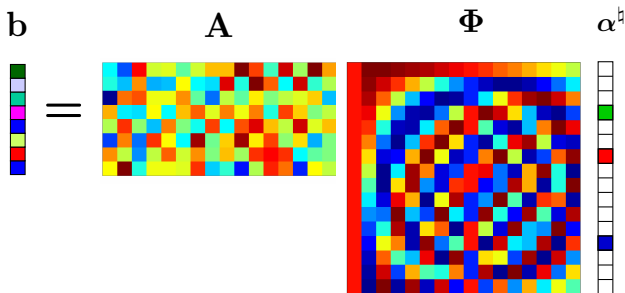


Sparse representations strike back!

$$\mathbf{b} = \mathbf{A} \mathbf{x}^{\text{b}}$$


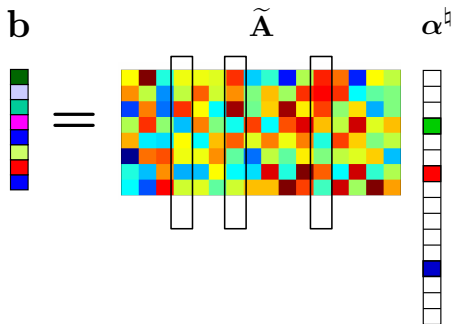
- ▶ $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $n < p$

Sparse representations strike back!



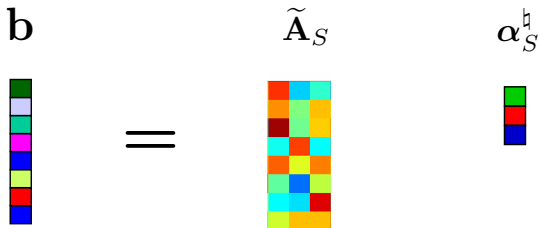
- ▶ $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $n < p$
- ▶ $\Psi \in \mathbb{R}^{p \times p}$, $\alpha^{\mathbf{b}} \in \mathbb{R}^p$, and $\|\alpha^{\mathbf{b}}\|_0 \leq s < n$

Sparse representations strike back!



- ▶ $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$, and $\|\hat{\boldsymbol{\alpha}}\|_0 \leq s < n < p$

Sparse representations strike back!

$$\mathbf{b} = \tilde{\mathbf{A}}_S \alpha_S^{\natural}$$


A fundamental impact:

The matrix $\tilde{\mathbf{A}}$ effectively becomes *overcomplete*.

We could easily solve for α^{\natural} (and hence \mathbf{x}^{\natural}) if we knew *the location of the non-zero entries of \mathbf{x}^{\natural}* .

Sparse recovery via the Lasso

Definition (Least absolute shrinkage and selection operator (Lasso))

$$\hat{\alpha}_{\text{lasso}} := \arg \min_{\alpha \in \mathbb{R}^p} \left\| \mathbf{b} - \tilde{\mathbf{A}}\alpha \right\|_2^2 + \rho \|\alpha\|_1$$

with some $\rho \geq 0$.

The second term in the objective function is called the *regularizer*.

Here ρ is called the *regularization parameter*. It is used to trade off the objectives:

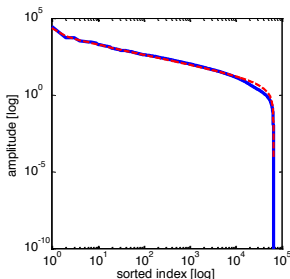
- ▶ Minimize $\|\mathbf{b} - \tilde{\mathbf{A}}\alpha\|_2^2$, so that the solution is consistent with the observations
- ▶ Minimize $\|\alpha\|_1$, so that the solution has the desired sparsity structure

The problem is efficiently solvable via *convex optimization*.

Compressible signals

Real signals may not be exactly sparse, but approximately sparse, or *compressible*.

Roughly speaking, a vector $\alpha := (\alpha_1, \dots, \alpha_p)^T \in \mathbb{R}^p$ is compressible if the number of its significant components, $|\{k : |\alpha_k| \geq t, 1 \leq k \leq p\}|$, is small.



▶ **Cameraman@MIT.**

- ▶ **Solid curve:** Sorted wavelet coefficients of the cameraman image.
- ▶ **Dashed curve:** Expected order statistics of generalized Pareto distribution with shape parameter 1.67.

Performance guarantees

Theorem (Existence of a stable solution in polynomial time [?])

This Lasso can be solved in polynomial time in terms of the inputs n and p . If the signal \mathbf{x}^{\natural} is s -sparse and the noise \mathbf{w} is Gaussian with variance σ^2 , then choosing

$\rho = \sqrt{\frac{16\sigma^2 \log p}{n}}$ yields an error of

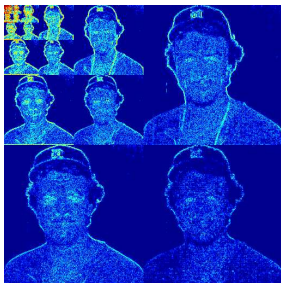
$$\|\hat{\boldsymbol{\alpha}}_{\text{lasso}} - \boldsymbol{\alpha}^{\natural}\|_2 \leq \frac{8\sigma}{\mu(\mathbf{A})} \sqrt{\frac{s \ln p}{n}},$$

with probability at least $1 - c_1 \exp(-c_2 n \rho^2)$, where c_1 and c_2 are absolute constants, and $\mu(\mathbf{A}) > 0$ encodes the difficulty of the problem.

- ▶ Hence, the number of measurements required is $\mathcal{O}(s \ln p)$ – this may be *much* smaller than p .
- ▶ $\mu(\mathbf{A})$ can be made large, for example, by letting the entries of \mathbf{A} be *i.i.d. Gaussian*

Structured sparsity

Real-world signals tend not to have arbitrary sparsity patterns, but instead tend to follow more specific patterns (e.g., clustering; tree structures)



Typical measurement matrix constructions (e.g., i.i.d. Gaussian) and decoding algorithms (e.g., basis pursuit) **do not exploit these more refined structures**

Subsampled measurement matrices

Subsampled measurement matrices

Subsampled measurement matrices take the form [?]

$$\mathbf{A} = \mathbf{P}_\Omega \Psi,$$

where:

- ▶ $\Psi \in \mathbb{C}^{p \times p}$: orthonormal/unitary *measurement basis* matrix
- ▶ $\mathbf{P}_\Omega : \mathbb{C}^p \rightarrow \mathbb{C}^n$: *subsampling matrix* such that $[\mathbf{P}_\Omega \mathbf{x}]_l = \mathbf{x}_{\Omega_l}$ (i.e., keep only the rows indexed by Ω)

Subsampled measurement matrices

Subsampled measurement matrices

Subsampled measurement matrices take the form [?]

$$\mathbf{A} = \mathbf{P}_\Omega \Psi,$$

where:

- ▶ $\Psi \in \mathbb{C}^{p \times p}$: orthonormal/unitary *measurement basis* matrix
- ▶ $\mathbf{P}_\Omega : \mathbb{C}^p \rightarrow \mathbb{C}^n$: *subsampling matrix* such that $[\mathbf{P}_\Omega \mathbf{x}]_l = \mathbf{x}_{\Omega_l}$ (i.e., keep only the rows indexed by Ω)

Motivation:

- ▶ Improved computational efficiency (e.g., Hadamard or Fourier Ψ)
- ▶ Applications where measurements must be in a certain basis (e.g., Fourier in MRI)

Fundamental question: **How do we choose the “best” Ω ?**

- ▶ This naturally leads to *discrete optimization* problems.

Learning-based compressive subsampling

The idea of the *learning-based compressive subsampling* (LB-CS) framework [?] is to learn Ω based on a set of *training signals*.

LB-CS: Problem statement

Given a set of m *training signals* $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{C}^p$, find an index set Ω of a given cardinality n such that a related *test signal* \mathbf{x} can reliably be recovered given the subsampled measurement vector $\mathbf{b} = \mathbf{P}_\Omega \Psi \mathbf{x}$.

A natural approach is to *choose the indices that have the highest energy* in some sense.

Optimizing the energy captured (I)

General optimization template

We study the following class of problems:

$$\hat{\Omega} = \arg \max_{\Omega : |\Omega|=n} F(\Omega),$$

where

$$F(\Omega) := f\left(\|\mathbf{P}_{\Omega} \Psi \mathbf{x}_1\|_2^2, \dots, \|\mathbf{P}_{\Omega} \Psi \mathbf{x}_m\|_2^2\right)$$

for some function f .

Observe that $\|\mathbf{P}_{\Omega} \Psi \mathbf{x}_j\|_2^2$ is precisely the energy in \mathbf{x}_j captured by the subsamples corresponding to Ω . It will be convenient to write

$$\|\mathbf{P}_{\Omega} \Psi \mathbf{x}_j\|_2^2 = \sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2,$$

where ψ_i is the i -th row of Ψ

Optimizing the energy captured (II)

Average energy

Using $f_{\text{avg}}(\alpha_1, \dots, \alpha_m) := \frac{1}{m} \sum_{j=1}^m \alpha_j$ yields

$$\hat{\Omega} = \arg \max_{\Omega: |\Omega|=n} \frac{1}{m} \sum_{j=1}^m \sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2,$$

Optimizing the energy captured (II)

Average energy

Using $f_{\text{avg}}(\alpha_1, \dots, \alpha_m) := \frac{1}{m} \sum_{j=1}^m \alpha_j$ yields

$$\hat{\Omega} = \arg \max_{\Omega: |\Omega|=n} \frac{1}{m} \sum_{j=1}^m \sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2,$$

Generalized average energy

Using $f_{\text{gen}}(\alpha_1, \dots, \alpha_m) := \frac{1}{m} \sum_{j=1}^m g(\alpha_j)$ yields

$$\hat{\Omega} = \arg \max_{\Omega: |\Omega|=n} \frac{1}{m} \sum_{j=1}^m g \left(\sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2 \right).$$

Here we let $g : [0, 1] \rightarrow \mathbb{R}$ be an increasing concave function with $g(0) = 0$.

Optimizing the energy captured (II)

Average energy

Using $f_{\text{avg}}(\alpha_1, \dots, \alpha_m) := \frac{1}{m} \sum_{j=1}^m \alpha_j$ yields

$$\hat{\Omega} = \arg \max_{\Omega: |\Omega|=n} \frac{1}{m} \sum_{j=1}^m \sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2,$$

Generalized average energy

Using $f_{\text{gen}}(\alpha_1, \dots, \alpha_m) := \frac{1}{m} \sum_{j=1}^m g(\alpha_j)$ yields

$$\hat{\Omega} = \arg \max_{\Omega: |\Omega|=n} \frac{1}{m} \sum_{j=1}^m g \left(\sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2 \right).$$

Here we let $g : [0, 1] \rightarrow \mathbb{R}$ be an increasing concave function with $g(0) = 0$.

Worst-case energy

Using $f_{\text{min}}(\alpha_1, \dots, \alpha_m) := \min_{j=1, \dots, m} \alpha_j$ yields

$$\hat{\Omega} = \arg \max_{\Omega: |\Omega|=n} \min_{j=1, \dots, m} \sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2.$$

Interpretations

The preceding choices of f can be interpreted as follows:

- ▶ $f = f_{\text{avg}}$: Choose the indices maximizing the *average energy* captured in the training signals $\mathbf{x}_1, \dots, \mathbf{x}_m$;
- ▶ $f = f_{\text{gen}}$: Instead of maximizing the average energy, maximize the *average of some suitably-designed function* $g(\cdot)$;
- ▶ $f = f_{\text{min}}$: Choose the indices maximizing the *worst-case energy* captured in the training signals $\mathbf{x}_1, \dots, \mathbf{x}_m$.

The worst-case criterion f_{min} may be preferable in some cases, but it tends to be less robust to “outliers”, e.g., compared to f_{avg} .

Linear decoding performance

Capturing energy sounds like a reasonable criterion, but does it actually correspond to good recovery performance? The answer is yes for a particular choice of decoder.

Linear decoder

We consider a linear decoder that expands \mathbf{b} to a p -dimensional vector by placing zeros in the entries corresponding to Ω^c , and then applies the adjoint $\Psi^* = \Psi^{-1}$:

$$\hat{\mathbf{x}} = \Psi^* \mathbf{P}_{\Omega}^T \mathbf{b}.$$

Exercise: Show that this decoder is equivalent to least-squares decoding. (Recall that the observations are given by $\mathbf{b} = \mathbf{P}_{\Omega} \Psi \mathbf{x}$)

Linear decoding performance

Capturing energy sounds like a reasonable criterion, but does it actually correspond to good recovery performance? The answer is yes for a particular choice of decoder.

Linear decoder

We consider a linear decoder that expands \mathbf{b} to a p -dimensional vector by placing zeros in the entries corresponding to Ω^c , and then applies the adjoint $\Psi^* = \Psi^{-1}$:

$$\hat{\mathbf{x}} = \Psi^* \mathbf{P}_\Omega^T \mathbf{b}.$$

Exercise: Show that this decoder is equivalent to least-squares decoding. (Recall that the observations are given by $\mathbf{b} = \mathbf{P}_\Omega \Psi \mathbf{x}$)

Theorem

The ℓ_2 estimation error of the above decoder is

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \|\mathbf{x}\|_2^2 - \|\mathbf{P}_\Omega \Psi \mathbf{x}\|_2^2.$$

Exercise: Prove this theorem. (Recall that Ψ is assumed to be unitary)

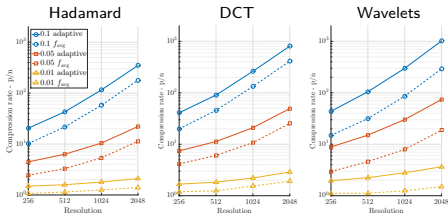
Note: This theorem shows that **maximizing the captured energy amounts to minimizing the error of the linear decoder.**

Example: Natural Images

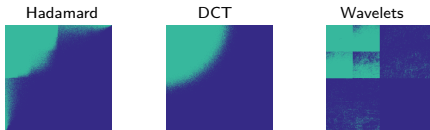
As an illustration, let's apply this approach to subsampling natural images in the Hadamard, discrete cosine transform (DCT), and wavelet domains.



The performance of learning-based (with f_{avg}) and *adaptive* (best possible samples *image-by-image*) subsampling:



The learned subsampling patterns:



Incorporating constraints

So far, we have considered simply selecting the best k indices $\Omega \subseteq \{1, \dots, p\}$. In some cases, we might want to impose *constraints* on Ω .

LB-CS with Constraints: Problem statement

Given a set of m *training signals* $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{C}^p$ and a *constraint set* \mathcal{A} (containing subsets of $\{1, \dots, p\}$ of cardinality n), find an index set $\Omega \in \mathcal{A}$ such that a related *test signal* \mathbf{x} can reliably be recovered given $\mathbf{b} = \mathbf{P}_\Omega \Psi \mathbf{x}$.

Incorporating constraints

So far, we have considered simply selecting the best k indices $\Omega \subseteq \{1, \dots, p\}$. In some cases, we might want to impose *constraints* on Ω .

LB-CS with Constraints: Problem statement

Given a set of m *training signals* $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{C}^p$ and a *constraint set* \mathcal{A} (containing subsets of $\{1, \dots, p\}$ of cardinality n), find an index set $\Omega \in \mathcal{A}$ such that a related *test signal* \mathbf{x} can reliably be recovered given $\mathbf{b} = \mathbf{P}_\Omega \Psi \mathbf{x}$.

Examples:

- ▶ **Multi-level sampling:** Split the indices into K disjoint groups G_1, \dots, G_K ; set

$$\mathcal{A} = \left\{ \Omega : \Omega \text{ contains } n_k \text{ indices from } G_k, \forall k = 1, \dots, K \right\}$$

where n_1, \dots, n_K are integers such that $\sum_{k=1}^K n_k = n$.

- ▶ For example, G_1 might contain the *lowest frequencies*, and G_K the *highest frequencies*.
- ▶ **Rooted-connected tree structure:** The indices form a rooted-connected subtree of the Wavelet tree [?].

Optimizing the energy captured with constraints

General optimization template with constraints

We study the following class of problems:

$$\hat{\Omega} = \arg \max_{\Omega \in \mathcal{A}} F(\Omega),$$

where

$$F(\Omega) := f\left(\|\mathbf{P}_{\Omega} \Psi \mathbf{x}_1\|_2^2, \dots, \|\mathbf{P}_{\Omega} \Psi \mathbf{x}_m\|_2^2\right)$$

for some function f .

Notes:

- ▶ Everything is the same as before, except we have included the constraint $\Omega \in \mathcal{A}$.
- ▶ The choices $f = f_{\text{avg}}, f_{\text{gen}}, f_{\text{min}}$ are still all suitable; however, the optimization problem may become more difficult depending on the choice of \mathcal{A} .
- ▶ For the two examples of \mathcal{A} on the previous slide, we can still efficiently solve exactly for f_{avg} , and approximately for f_{gen} [?].

Preview of results

Later in the course, we will explore the following theoretical results:

- ▶ **Discrete optimization guarantees:**
 - ▶ The optimization of f_{avg} can be solved exactly by exploiting *modularity*
 - ▶ The optimization of f_{gen} can be solved approximately by exploiting *submodularity*
 - ▶ The optimization of f_{min} can be solved approximately via the *robust submodular optimization* framework
- ▶ **Statistical generalization bound:** If the training images and test signal are drawn independently from a common *unknown* distribution, then under the f_{avg} criterion, the average energy captured in the test signal is within

$$\sqrt{\frac{2}{m} \left(\log |\mathcal{A}| + \log \frac{2}{\eta} \right)}$$

of the *best possible* – note that this approaches zero as m increases. This result is proved via an *empirical risk minimization* perspective.

Part II: Introduction to Submodularity

Recommended reading:

- ▶ *Submodular function maximization*, Krause and Golovin, 2012
- ▶ *Submodular functions and convexity*, Lovász, 1983

Definition of Submodularity

Let $V = \{1, \dots, n\}$ be the *ground set*, and let 2^V be the set of all subsets of V . For a function $f : 2^V \rightarrow \mathbb{R}$, set $S \subset V$, and element $e \in V \setminus S$, define the *discrete derivative/difference*

$$\Delta(e|S) = f(S \cup \{e\}) - f(S)$$

Definition (Submodularity)

A function $f : 2^V \rightarrow \mathbb{R}$ is said to be:

- ▶ *submodular* if, for all $S \subseteq T \subseteq V$ and $e \in V \setminus \{e\}$, it holds $\Delta(e|S) \geq \Delta(e|T)$;
- ▶ *modular* if it always holds that $\Delta(e|S) = \Delta(e|T)$;
- ▶ *supermodular* if it always holds that $\Delta(e|S) \leq \Delta(e|T)$.

The Intuition: “Diminishing returns” – adding to a smaller set gains you more than adding to a larger set.

Definition of Submodularity

Let $V = \{1, \dots, n\}$ be the *ground set*, and let 2^V be the set of all subsets of V . For a function $f : 2^V \rightarrow \mathbb{R}$, set $S \subset V$, and element $e \in V \setminus S$, define the *discrete derivative/difference*

$$\Delta(e|S) = f(S \cup \{e\}) - f(S)$$

Definition (Submodularity)

A function $f : 2^V \rightarrow \mathbb{R}$ is said to be:

- ▶ *submodular* if, for all $S \subseteq T \subseteq V$ and $e \in V \setminus \{e\}$, it holds $\Delta(e|S) \geq \Delta(e|T)$;
- ▶ *modular* if it always holds that $\Delta(e|S) = \Delta(e|T)$;
- ▶ *supermodular* if it always holds that $\Delta(e|S) \leq \Delta(e|T)$.

The Intuition: “Diminishing returns” – adding to a smaller set gains you more than adding to a larger set.

Definition (Monotonicity)

A function $f : 2^V \rightarrow \mathbb{R}$ is said to be *monotone* if, for all $S \subseteq T \subseteq V$, it holds that $f(S) \leq f(T)$.

Equivalent definitions of submodularity

Theorem

For any function $f : 2^V \rightarrow \mathbb{R}$, all of the following are equivalent:

- ▶ $\Delta(e|S) \geq \Delta(e|T)$ for all $S \subseteq T$, e
- ▶ $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$ for all S, T
- ▶ $\Delta(e|S) \geq \Delta(e|A \cup \{e'\})$ for all S, e, e'
- ▶ $f(T) \leq f(S) + \sum_{e \in T \setminus S} \Delta(e|S)$ for all $S \subseteq T$

See [?] for more equivalent definitions.

Equivalent definitions of submodularity

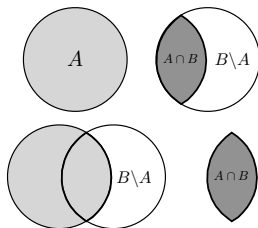
Theorem

For any function $f : 2^V \rightarrow \mathbb{R}$, all of the following are equivalent:

- ▶ $\Delta(e|S) \geq \Delta(e|T)$ for all $S \subseteq T$, e
- ▶ $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$ for all S, T
- ▶ $\Delta(e|S) \geq \Delta(e|A \cup \{e'\})$ for all S, e, e'
- ▶ $f(T) \leq f(S) + \sum_{e \in T \setminus S} \Delta(e|S)$ for all $S \subseteq T$

See [?] for more equivalent definitions.

A visual description of the second of these:



Relations to convexity and concavity

Submodular functions share some properties of concave functions:

- ▶ Diminishing returns
- ▶ Any local maximum within a factor of $\frac{1}{2}$ of globally optimal

Relations to convexity and concavity

Submodular functions share some properties of concave functions:

- ▶ Diminishing returns
- ▶ Any local maximum within a factor of $\frac{1}{2}$ of globally optimal

...but they also share some properties with convex functions:

- ▶ Unconstrained minimization is “easy” (though constrained minimization is extremely hard!)
- ▶ The Lovász extension [?] is convex

Relations to convexity and concavity

Submodular functions share some properties of concave functions:

- ▶ Diminishing returns
- ▶ Any local maximum within a factor of $\frac{1}{2}$ of globally optimal

...but they also share some properties with convex functions:

- ▶ Unconstrained minimization is “easy” (though constrained minimization is extremely hard!)
- ▶ The Lovász extension [?] is convex

They also fail to share some of the nicest properties of each; in particular, the maximum or minimum of two submodular functions is not submodular in general.

Properties of submodular functions

Properties of submodular functions

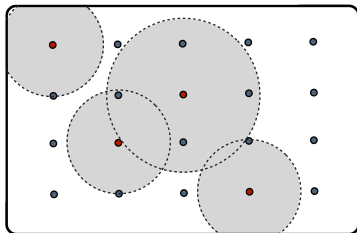
Let f_1 and f_2 be submodular functions.

1. **Linear combinations:** If c_1, c_2 are positive, then $f(S) = c_1 f_1(S) + c_2 f_2(S)$ is submodular.
2. **Concave of modular:** If $g : 2^V \rightarrow \mathbb{R}$ is modular and $h : \mathbb{R} \rightarrow \mathbb{R}$ is concave, then $f(S) = h(g(S))$ is submodular.
3. **Residual:** $f(S) = f_1(S \cup B) - f_1(B)$ is submodular for any B .
4. **Conditioning:** $f(S) = f_1(S \cap A)$ is submodular for any A .
5. **Reflection:** $f(S) = f_1(V \setminus S)$ is submodular.
6. **Truncation:** $f(S) = \min\{c, f(S)\}$ is submodular for any $c \in \mathbb{R}$.
7. **Minimum:** Although $f(S) = \min\{f_1(S), f_2(S)\}$ is not submodular in general, it is submodular when either $f_1 - f_2$ or $f_2 - f_1$ is monotone.

Examples of submodular functions (I)

Example 1: Let \mathbf{X} be a matrix, let V be the indices of its columns, and let \mathbf{X}_S be the submatrix formed by keeping only the columns indexed by S . Then $r(S) = \text{rank}(\mathbf{X}_S)$ is *monotone submodular*.

Example 2: Coverage functions are *monotone submodular*.



- Activated
- Deactivated

$f(S) = \text{Area covered by activating all sensors in } S$

Examples of submodular functions (II)

Example 3: Some examples from graph theory:

- ▶ For a bipartite graph $\{A, B\}$, let $S \subset A$. The the number of neighbors $\Gamma(S)$ of S is a *monotone submodular* function.
- ▶ For an undirected graph $G = (V, E)$, let $S \subset E$. Then the number of connected components $c(S)$ of S is *supermodular*.
- ▶ For an undirected graph $G = (V, E)$, associate with each edge e a capacity $c(e) \geq 0$. For $S \subset E$, let δS be the number of edges in E with exactly one end in S . Then $f(S) = \sum_{e \in \delta S} c(e)$ is *submodular*.

Examples of submodular functions (III)

Information-theoretic definitions

Given a joint PMF P_{XY} with marginals P_X , P_Y , and $P_{Y|X}$, define

$$H(X) = \sum_x P_X(x) \log \frac{1}{P_X(x)},$$

$$H(Y|X) = \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{Y|X}(y|x)} = \sum_x P_X(x) H(Y|X=x),$$

$$I(X; Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = H(Y) - H(Y|X) = H(X) - H(X|Y).$$

Example 4: Some examples from information theory:

- ▶ Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector, and let $X_S = \{X_i\}_{i \in S}$. Then the entropy $f(S) = H(\mathbf{X}_S)$ is *monotone submodular*.¹
- ▶ Let $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$ be (possibly dependent) random vectors. If the X_i are conditionally independent given \mathbf{Y} , then the mutual information $I(\mathbf{X}_S; \mathbf{Y})$ is *monotone submodular*.

¹For continuous random variables (with the above summations replaced by integrals), monotonicity may not hold, but the submodularity property remains.

Examples of submodular functions (IV)

Example 5: Some examples from learning-based CS:

- ▶ Our index selection rule for the *average* energy criterion was

$$f(\Omega) = \frac{1}{m} \sum_{j=1}^m \sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2.$$

This is a *modular* function.

- ▶ Our index selection rule for the *generalized average* energy criterion was

$$f(\Omega) = \frac{1}{m} \sum_{j=1}^m g\left(\sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2\right).$$

This is a *monotone submodular function* whenever $g(\cdot)$ is concave and increasing.

- ▶ Our index selection rule for the *worst-case* energy criterion was

$$f(\Omega) = \min_{j=1, \dots, m} \sum_{i \in \Omega} |\langle \psi_i, \mathbf{x}_j \rangle|^2.$$

This is *not submodular* in general, but it falls under the setting of *robust submodular optimization* [?], which we will study later.

References I

- [1] Luca Baldassarre, Yen-Huan Li, Jonathan Scarlett, Baran Gözcü, Ilija Bogunovic, and Volkan Cevher.
Learning-based compressive subsampling.
<http://arxiv.org/abs/1510.06188>.
- [2] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde.
Model-based compressive sensing.
Information Theory, IEEE Transactions on, 56(4):1982–2001, 2010.
- [3] Simon Foucart and Holger Rauhut.
A mathematical introduction to compressive sensing.
Springer, 2013.
- [4] Andreas Krause, H. Brendan McMahan, Carlos Guestrin, and Anupam Gupta.
Robust submodular observation selection.
Journal of Machine Learning Research, pages 2761–2801, Dec. 2008.
- [5] L. Lovász.
Submodular functions and convexity.
In Mathematical Programming The State of the Art, pages 235–257. Springer, 1983.

References II

- [6] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu.
A unified framework for high-dimensional analysis of M -estimators with decomposable regularizers.
Stat. Sci., 27(4):538–557, 2012.
- [7] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher.
An analysis of approximations for maximizing submodular set functions – i.
Mathematical Programming, 14(1):265–294, 1978.