

Submodular Functions: from Discrete to Continuous Domains

Francis Bach

INRIA - Ecole Normale Supérieure



ETH Zürich - April 2016

Submodular functions

From discrete to continuous domains

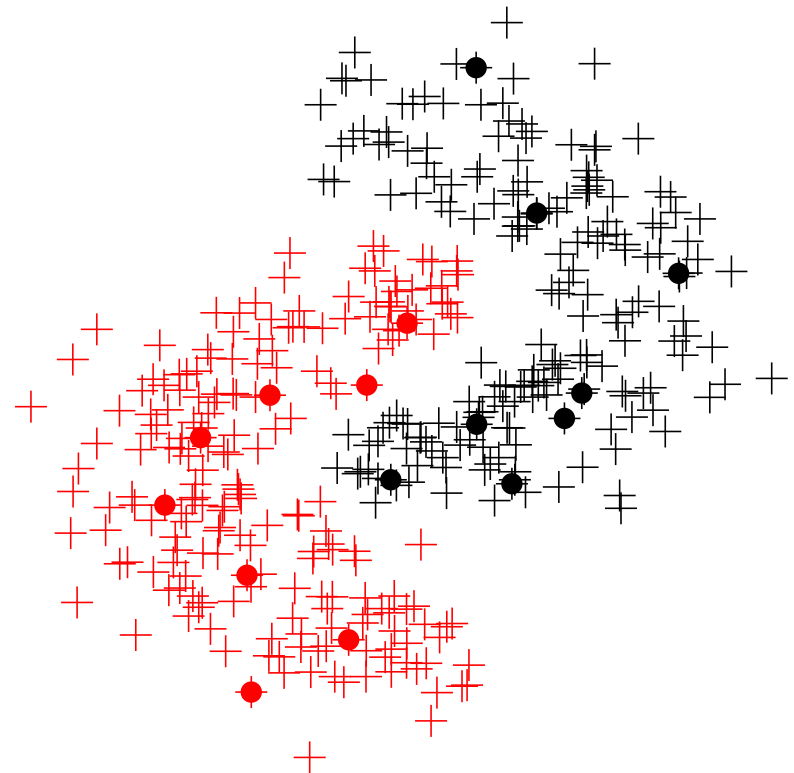
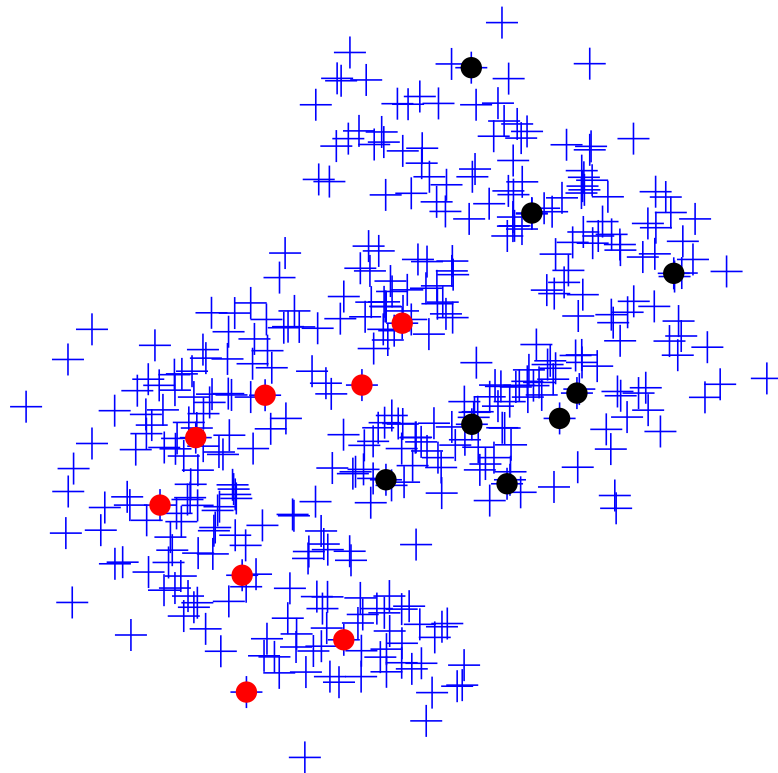
Summary

- **Which functions can be minimized in polynomial time?**
 - Beyond convex functions
- **Submodular functions**
 - Not convex, ... but “equivalent” to convex functions
 - Usually defined on $\{0, 1\}^n$
 - Extension to continuous domains
- **Preprint available on ArXiv, second version (Bach, 2015)**

Submodularity (almost) everywhere

Clustering

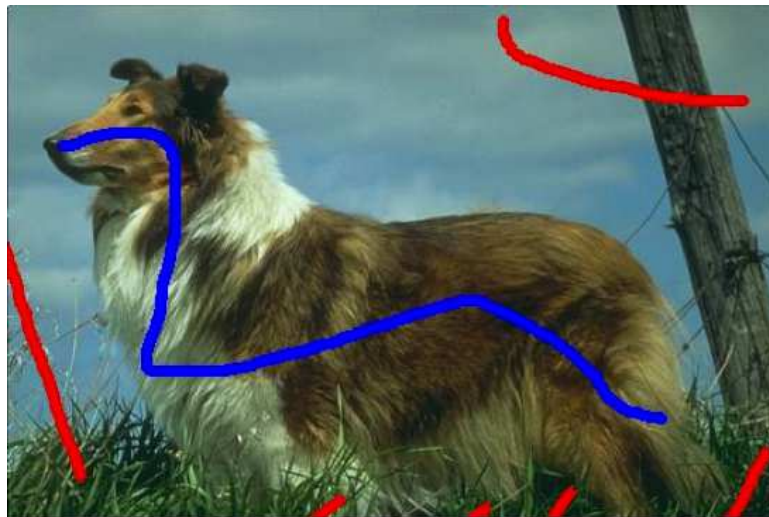
- Semi-supervised clustering



- Submodular function minimization

Submodularity (almost) everywhere

Graph cuts and image segmentation

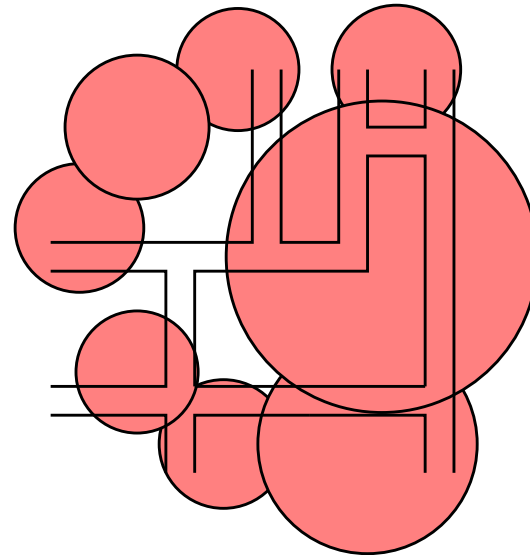
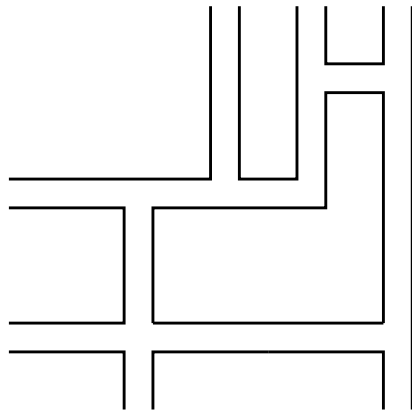


- Submodular function minimization

Submodularity (almost) everywhere

Sensor placement

- Each sensor covers a certain area (Krause and Guestrin, 2005)
 - Goal: maximize coverage



- Submodular function maximization
- Extension to experimental design (Seeger, 2009)

Submodularity (almost) everywhere

Image denoising

- Total variation denoising (Chambolle, 2005)



- Submodular convex optimization problem

Submodularity (almost) everywhere

Combinatorial optimization problems

- Set $V = \{1, \dots, n\}$
- Power set $2^V =$ set of all subsets, of cardinality 2^n
- Minimization/maximization of a set-function $F : 2^V \rightarrow \mathbb{R}$.

$$\min_{A \subset V} F(A) = \min_{A \in 2^V} F(A)$$

Submodularity (almost) everywhere

Combinatorial optimization problems

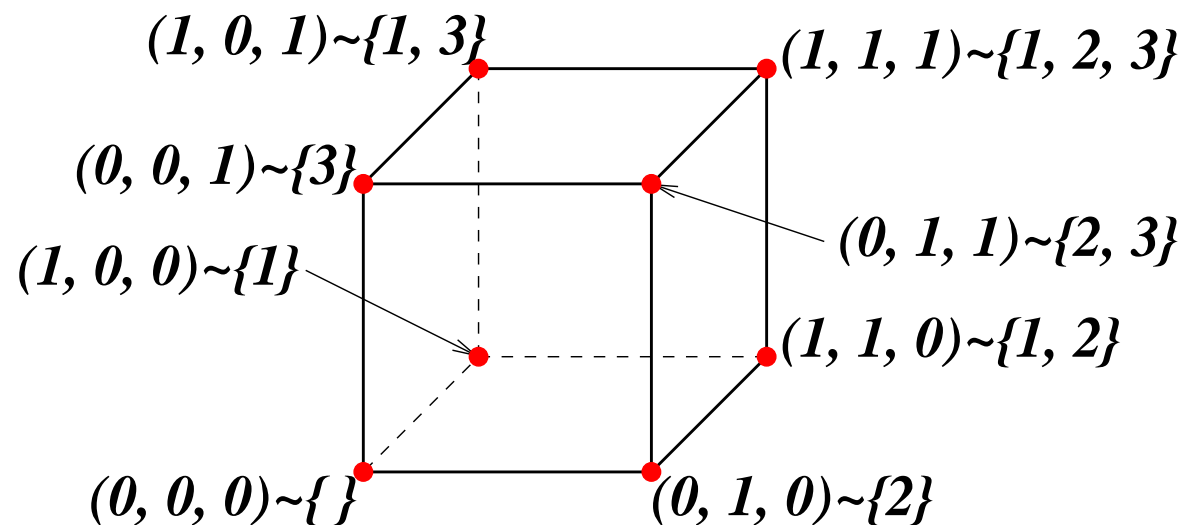
- Set $V = \{1, \dots, n\}$
- Power set $2^V =$ set of all subsets, of cardinality 2^n
- Minimization/maximization of a set-function $F : 2^V \rightarrow \mathbb{R}$.

$$\min_{A \subset V} F(A) = \min_{A \in 2^V} F(A)$$

- Reformulation as (pseudo) Boolean function

$$\min_{x \in \{0,1\}^n} H(x)$$

with $H : \{0, 1\}^n \rightarrow \mathbb{R}$
and $\forall A \subset V, H(1_A) = F(A)$



Outline

1. Submodular set-functions

- Definitions, examples
- Links with convexity through Lovász extension
- Minimization by convex optimization

2. From discrete to continuous domains

- Nonpositive second-order derivatives
- Invariances and examples
- Extensions on product measures through optimal transport

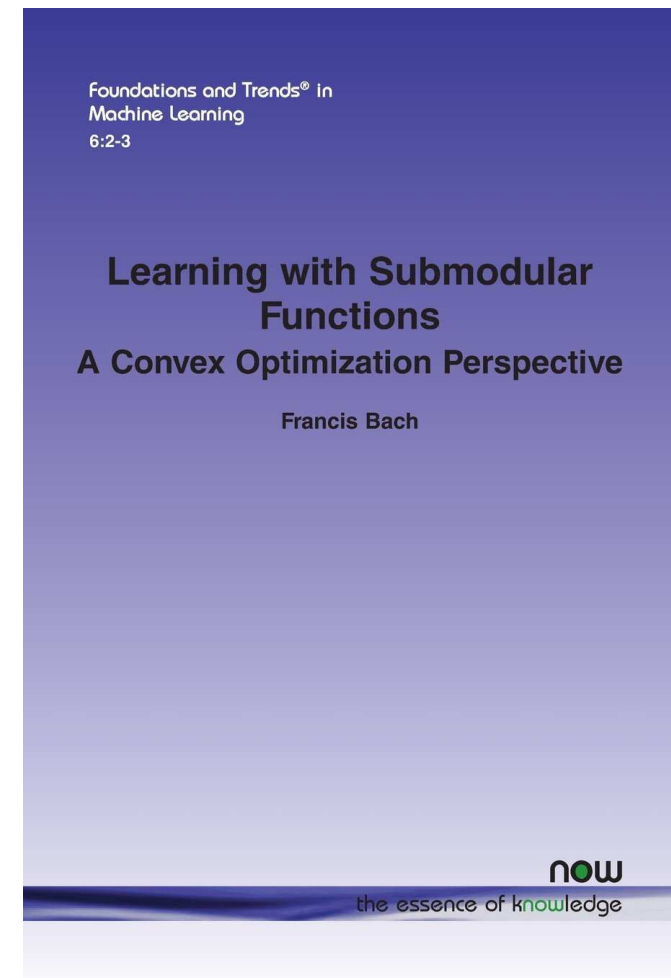
3. Minimization of continuous submodular functions

- Subgradient descent
- Frank-Wolfe optimization

Submodular functions - References

- Reference book based on combinatorial optimization
 - *Submodular Functions and Optimization* (Fujishige, 2005)

- Tutorial monograph based on convex optimization (Bach, 2013)
 - *Learning with submodular functions: a convex optimization perspective*



Submodular functions

Definitions

- **Definition:** $H : \{0, 1\}^n \rightarrow \mathbb{R}$ is **submodular** if and only if

$$\forall x, y \in \{0, 1\}^n, \quad H(x) + H(y) \geq H(\max\{x, y\}) + H(\min\{x, y\})$$

- NB: equality for *modular* functions (linear functions of x)
- Always assume $H(0) = 0$

Submodular functions

Definitions

- **Definition:** $H : \{0, 1\}^n \rightarrow \mathbb{R}$ is **submodular** if and only if

$$\forall x, y \in \{0, 1\}^n, \quad H(x) + H(y) \geq H(\max\{x, y\}) + H(\min\{x, y\})$$

- NB: equality for *modular* functions (linear functions of x)
- Always assume $H(0) = 0$
- **Equivalent definition:** (with $e_i \in \mathbb{R}^n$ i -th canonical basis vector)

$$\forall i \in \{1, \dots, n\}, \quad x \mapsto H(x + e_i) - H(x) \text{ is non-increasing}$$

- “**Concave property**”: Diminishing returns

Submodular functions - Examples

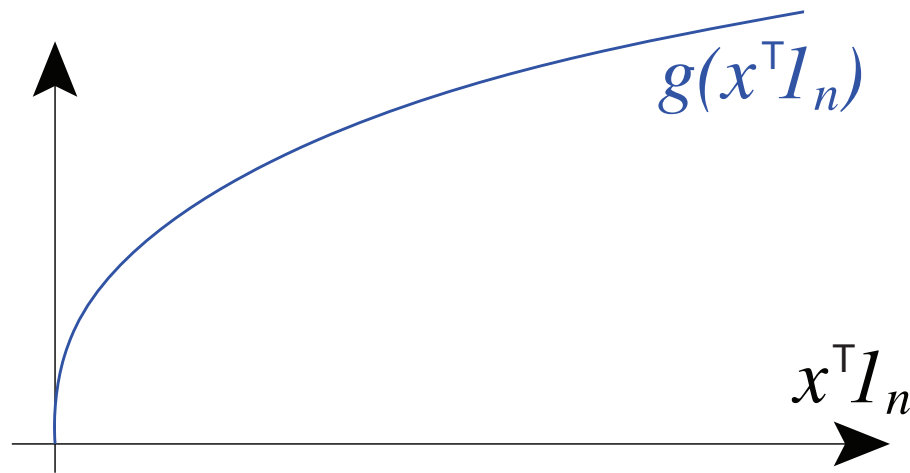
(see, e.g., Fujishige, 2005; Bach, 2013)

- Concave functions of the cardinality
- Cuts
- Entropies
 - Joint entropy of $(X_k)_{x_k=1}$, from n random variables X_1, \dots, X_n
- Functions of eigenvalues of sub-matrices
- Network flows
- Rank functions of matroids

Examples of submodular functions

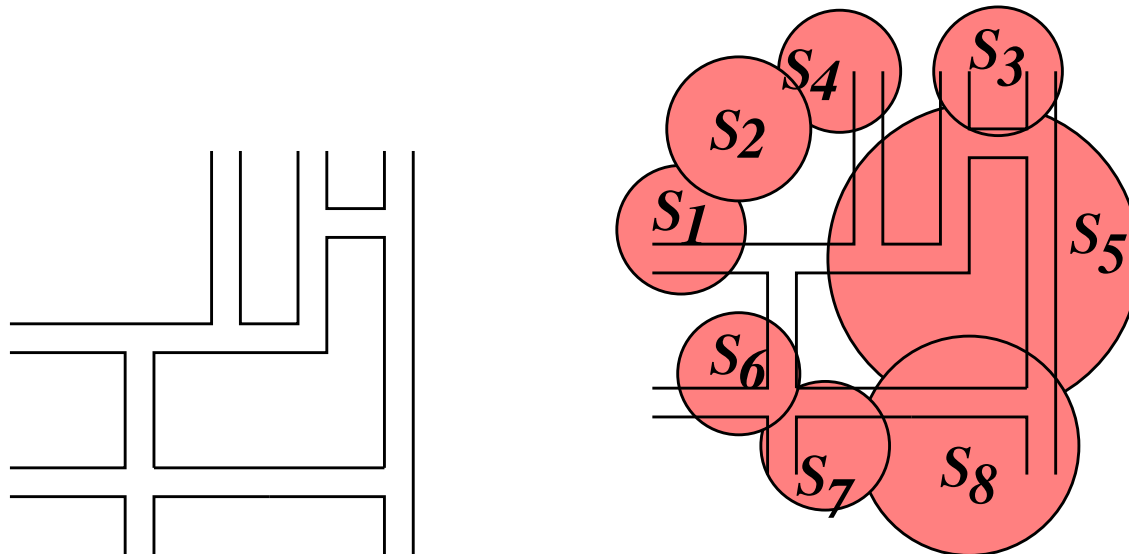
Cardinality-based functions

- **Modular function:** $H(x) = w^\top x$ for $w \in \mathbb{R}^n$
 - Cardinality example: If $w = 1_n$, then $H(x) = 1_n^\top x$
- If g is a **concave function**, then $H : x \mapsto g(1_n^\top x)$ is submodular
 - Diminishing return property



Examples of submodular functions

Covers

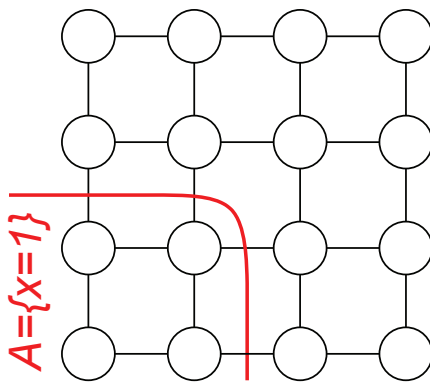


- Let W be any “base” set, and for each $k \in V$, a set $S_k \subset W$
- Set cover defined as $H(x) = \left| \bigcup_{x_k=1} S_k \right|$

Examples of submodular functions

Cuts

- Given a (un)directed graph, with vertex set $V = \{1, \dots, n\}$ and edge set $E \subset V \times V$
 - $H(x)$ is the total number of edges going from $\{x = 1\}$ to $\{x = 0\}$.

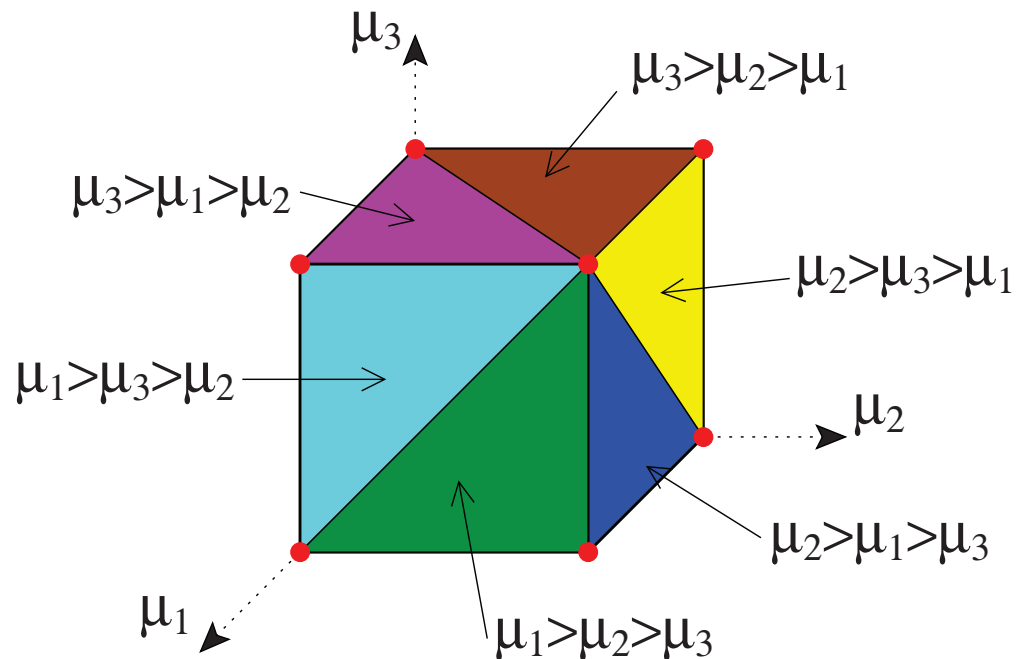


- Generalization with $d : \{1, \dots, n\} \times \{1, \dots, n\} \rightarrow \mathbb{R}_+$

$$H(x) = \sum_{j,k} d(k, j)(x_k - x_j)_+$$

Choquet integral (Choquet, 1954) - Lovász extension

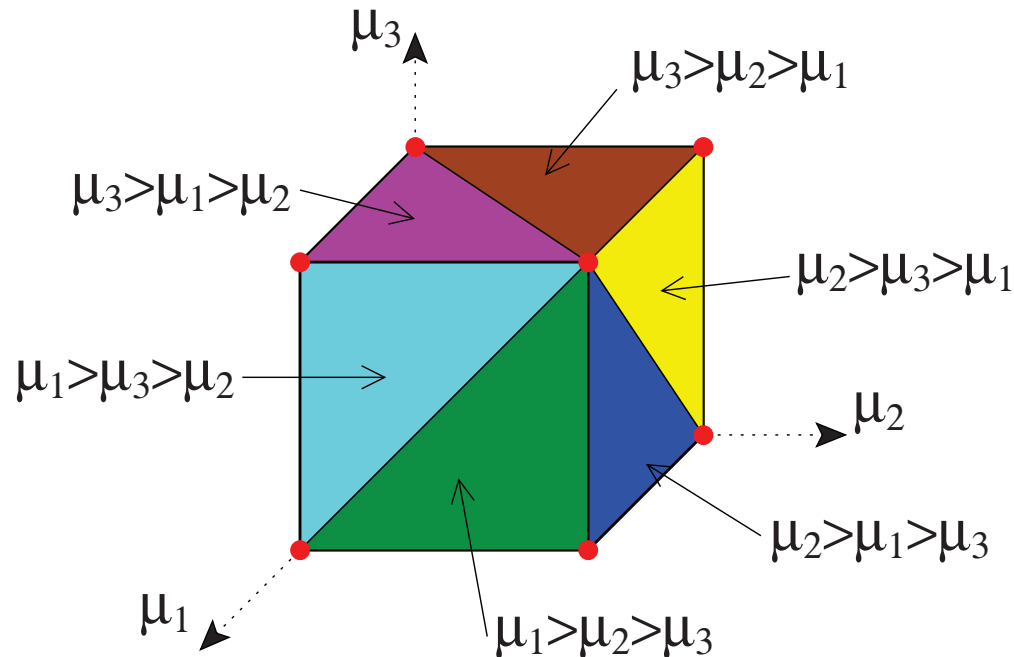
- Subsets may be identified with elements of $\{0, 1\}^n$
- Given **any** function H and $\mu \in \mathbb{R}^n$ such that $\mu_{j_1} \geq \dots \geq \mu_{j_n}$



Choquet integral (Choquet, 1954) - Lovász extension

- Subsets may be identified with elements of $\{0, 1\}^n$
- Given **any** function H and $\mu \in \mathbb{R}^n$ such that $\mu_{j_1} \geq \dots \geq \mu_{j_n}$, define:

$$h(\mu) = \sum_{k=1}^n \mu_{j_k} [H(e_{j_1} + \dots + e_{j_k}) - H(e_{j_1} + \dots + e_{j_{k-1}})]$$



Choquet integral (Choquet, 1954) - Lovász extension

- Subsets may be identified with elements of $\{0, 1\}^n$
- Given **any** function H and $\mu \in \mathbb{R}^n$ such that $\mu_{j_1} \geq \dots \geq \mu_{j_n}$, define:

$$h(\mu) = \sum_{k=1}^n \mu_{j_k} [H(e_{j_1} + \dots + e_{j_k}) - H(e_{j_1} + \dots + e_{j_{k-1}})]$$

- For $H(x) = w^\top x$, then $h(\mu) = w^\top \mu$
- For cuts, $h(\mu) = \sum_{k,j \in V} d(k,j) |\mu_k - \mu_j|$ is the *total variation*

Choquet integral (Choquet, 1954) - Lovász extension

- Subsets may be identified with elements of $\{0, 1\}^n$
- Given **any** function H and $\mu \in \mathbb{R}^n$ such that $\mu_{j_1} \geq \dots \geq \mu_{j_n}$, define:

$$h(\mu) = \sum_{k=1}^n \mu_{j_k} [H(e_{j_1} + \dots + e_{j_k}) - H(e_{j_1} + \dots + e_{j_{k-1}})]$$

- For $H(x) = w^\top x$, then $h(\mu) = w^\top \mu$
- For cuts, $h(\mu) = \sum_{k,j \in V} d(k,j) |\mu_k - \mu_j|$ is the *total variation*
- For any set-function H (even not submodular)
 - h is piecewise-linear and positively homogeneous
 - If $x \in \{0, 1\}^n$, $h(x) = H(x) \Rightarrow$ **extension from $\{0, 1\}^n$ to $[0, 1]^n$**

Submodular set-functions

Links with convexity (Lovász, 1982)

1. H is submodular if and only if h is convex

2. If H is submodular, then

$$\min_{x \in \{0,1\}^n} H(x) = \min_{\mu \in \{0,1\}^n} h(\mu) = \min_{\mu \in [0,1]^n} h(\mu)$$

3. If H is submodular, then a **subgradient** of h at any μ may be computed by the “greedy algorithm”

- Order the components of $\mu \in \mathbb{R}^n$ as $\mu_{j_1} \geq \dots \geq \mu_{j_n}$
- Define $w_{j_k} = H(e_{j_1} + \dots + e_{j_k}) - H(e_{j_1} + \dots + e_{j_{k-1}})$ for all k
- Moreover $h(\mu) = w^\top \mu$

Submodular set-functions

Links with convexity (Lovász, 1982)

1. H is submodular if and only if h is convex

2. If H is submodular, then

$$\min_{x \in \{0,1\}^n} H(x) = \min_{\mu \in \{0,1\}^n} h(\mu) = \min_{\mu \in [0,1]^n} h(\mu)$$

3. If H is submodular, then a **subgradient** of h at any μ may be computed by the “greedy algorithm”

- **Consequences**

- **Submodular function minimization may be done in polynomial time**
- Ellipsoid algorithm in $O(n^5)$ (Grötschel et al., 1981)

Exact submodular function minimization

Combinatorial algorithms

- Algorithms based on $\min_{\mu \in [0,1]^n} h(\mu)$ and its dual problem
- Output the subset A and a dual **certificate of optimality**
- Best algorithms have **polynomial complexity** (Schrijver, 2000; Iwata et al., 2001; Orlin, 2009)
 - Typically $O(n^6)$ or more
- **Not practical for large problems...**

Submodular function minimization

Through convex optimization

- Convex non-smooth optimization problem

$$\min_{x \in \{0,1\}^n} H(x) = \min_{\mu \in \{0,1\}^n} h(\mu) = \min_{\mu \in [0,1]^n} h(\mu)$$

- **Important properties of h for convex optimization**
 - Polyhedral function
 - Known subgradients obtained from greedy algorithm
- **Generic algorithms** (blind to submodular structure)
 - Some with complexity bounds, some without
 - Subgradient, Frank-Wolfe, simplex, cutting-plane (ACCPM)
 - See Bach (2013) for details

Outline

1. Submodular set-functions

- Definitions, examples
- Links with convexity through Lovász extension
- Minimization by convex optimization

2. From discrete to continuous domains

- Nonpositive second-order derivatives
- Invariances and examples
- Extensions on product measures through optimal transport

3. Minimization of continuous submodular functions

- Subgradient descent
- Frank-Wolfe optimization

From discrete to continuous domains

- Main insight: $\{0, 1\}$ is totally ordered!

From discrete to continuous domains

- Main insight: $\{0, 1\}$ is totally ordered!
- Extension to $\{0, \dots, k - 1\}$: $H : \{0, \dots, k - 1\}^n \rightarrow \mathbb{R}$

$$\forall x, y, \quad H(x) + H(y) \geq H(\min\{x, y\}) + H(\max\{x, y\})$$

- Equivalent definition: with $(e_i)_{i \in \{1, \dots, n\}}$ canonical basis of \mathbb{R}^n

$$\forall x, i \neq j, \quad H(x + e_i) + H(x + e_j) \geq H(x) + H(x + e_i + e_j)$$

- See Lorentz (1953); Topkis (1978)

From discrete to continuous domains

- Main insight: $\{0, 1\}$ is totally ordered!
- Extension to $\{0, \dots, k-1\}$: $H : \{0, \dots, k-1\}^n \rightarrow \mathbb{R}$

$$\forall x, y, \quad H(x) + H(y) \geq H(\min\{x, y\}) + H(\max\{x, y\})$$

- Equivalent definition: with $(e_i)_{i \in \{1, \dots, n\}}$ canonical basis of \mathbb{R}^n

$$\forall x, i \neq j, \quad H(x + e_i) + H(x + e_j) \geq H(x) + H(x + e_i + e_j)$$

- See Lorentz (1953); Topkis (1978)

- Taylor expansion:

$$- H(x + e_i) + H(x + e_j) \approx 2H(x) + \frac{\partial H}{\partial x_i} + \frac{\partial H}{\partial x_j} + \frac{1}{2} \frac{\partial^2 H}{\partial x_i^2} + \frac{1}{2} \frac{\partial^2 H}{\partial x_j^2}$$

$$- H(x) + H(x + e_i + e_j) = 2H(x) + \frac{\partial H}{\partial x_i} + \frac{\partial H}{\partial x_j} + \frac{1}{2} \frac{\partial^2 H}{\partial x_i^2} + \frac{1}{2} \frac{\partial^2 H}{\partial x_j^2} + \frac{\partial^2 H}{\partial x_i \partial x_j}$$

From discrete to continuous domains

- **Main insight:** $\{0, 1\}$ is **totally ordered!**

- **Extension to** $\{0, \dots, k - 1\}$: $H : \{0, \dots, k - 1\}^n \rightarrow \mathbb{R}$

$$\forall x, y, \quad H(x) + H(y) \geq H(\min\{x, y\}) + H(\max\{x, y\})$$

- Equivalent definition: with $(e_i)_{i \in \{1, \dots, n\}}$ canonical basis of \mathbb{R}^n

$$\forall x, i \neq j, \quad H(x + e_i) + H(x + e_j) \geq H(x) + H(x + e_i + e_j)$$

- See Lorentz (1953); Topkis (1978)

- **Generalization to all totally ordered sets:** $\mathcal{X}_i \subset \mathbb{R}$

intervals + H twice differentiable: $\forall x \in \prod_{i=1}^n \mathcal{X}_i, \quad \frac{\partial^2 H}{\partial x_i \partial x_j}(x) \leq 0$

A “new” class of continuous functions

- Assume each $\mathcal{X}_i \subset \mathbb{R}$ is a compact interval, and (for simplicity) H twice differentiable:

$$\text{Submodularity} : \forall x \in \prod_{i=1}^n \mathcal{X}_i, \quad \frac{\partial^2 H}{\partial x_i \partial x_j}(x) \leq 0$$

- **Invariance** by
 - individual increasing smooth change of variables $H(\varphi_1(x_1), \dots, \varphi_n(x_n))$
 - adding arbitrary (smooth) separable functions $\sum_{i=1}^n v_i(x_i)$

A “new” class of continuous functions

- Assume each $\mathcal{X}_i \subset \mathbb{R}$ is a compact interval, and (for simplicity) H twice differentiable:

$$\text{Submodularity} : \forall x \in \prod_{i=1}^n \mathcal{X}_i, \quad \frac{\partial^2 H}{\partial x_i \partial x_j}(x) \leq 0$$

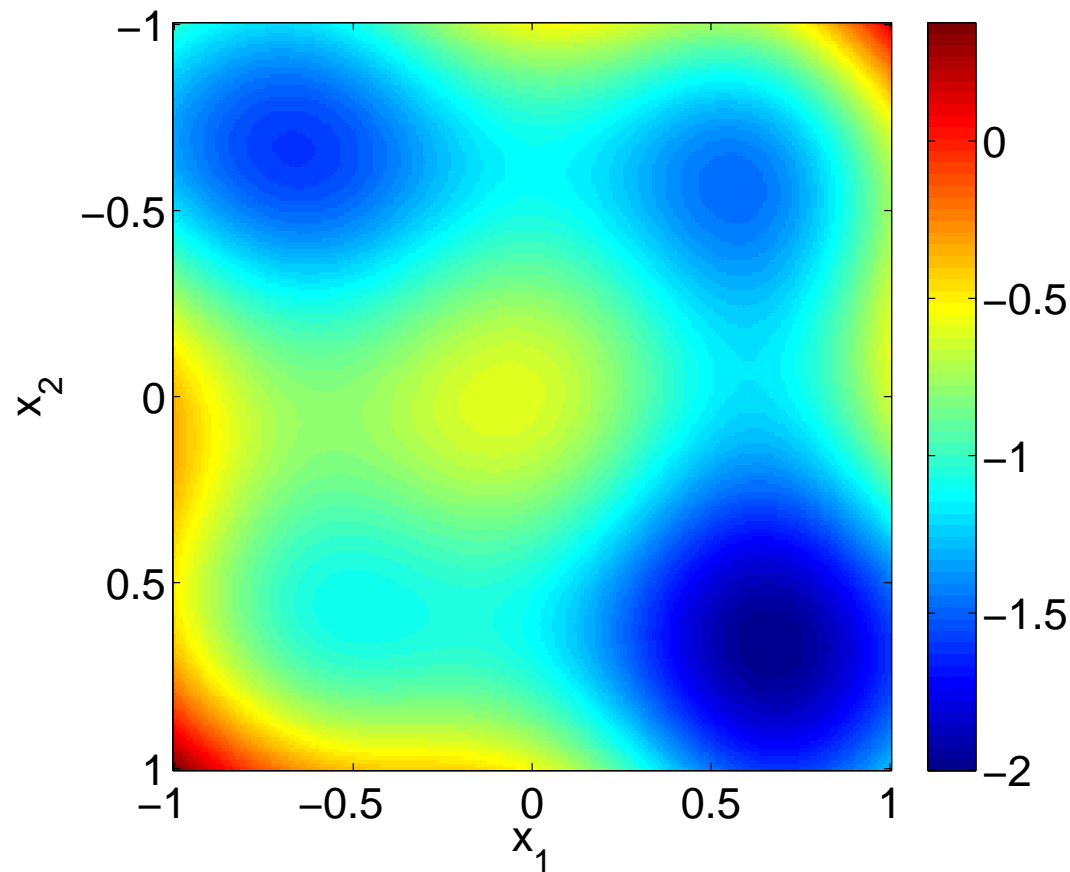
- **Invariance** by

- individual increasing smooth change of variables $H(\varphi_1(x_1), \dots, \varphi_n(x_n))$
- adding arbitrary (smooth) separable functions $\sum_{i=1}^n v_i(x_i)$

- **Examples**

- Quadratic functions with Hessians with non-negative off-diagonal entries (Kim and Kojima, 2003)
- $\varphi(x_i - x_j)$, φ convex; $\varphi(x_1 + \dots + x_n)$, φ concave; log det, etc...
- Monotone of order two (Carlier, 2003), Spence-Mirrlees condition (Milgrom and Shannon, 1994)

A “new” class of continuous functions



- Level sets of the submodular function $(x_1, x_2) \mapsto \frac{7}{20}(x_1 - x_2)^2 - e^{-4(x_1 - \frac{2}{3})^2} - \frac{3}{5}e^{-4(x_1 + \frac{2}{3})^2} - e^{-4(x_2 - \frac{2}{3})^2} - e^{-4(x_2 + \frac{2}{3})^2}$, with several local minima, local maxima and saddle points

Extensions to the space of product measures

- **Set-function:** $\mathcal{X}_i = \{0, 1\}$

- $[0, 1] \approx$ set of probability distributions on $\{0, 1\}$: $\mu_i = \mathbb{P}(X_i = 1)$
- Lovász extension: for $\mu \in [0, 1]^n$ such that $\mu_{j_1} \geq \dots \geq \mu_{j_n}$

$$\begin{aligned} h(\mu) &= \sum_{k=1}^n \mu_{j_k} [H(e_{j_1} + \dots + e_{j_k}) - H(e_{j_1} + \dots + e_{j_{k-1}})] \\ &= (1 - \mu_{j_1})H(0) + \sum_{k=1}^{n-1} (\mu_{j_k} - \mu_{j_{k+1}})H(e_{j_1} + \dots + e_{j_k}) + \mu_{j_n}H(1_n) \\ &= \mathbb{E}[H(1_{\mu_1 \geq t}, \dots, 1_{\mu_n \geq t})] \text{ for } t \text{ uniform in } [0, 1] \end{aligned}$$

$$\left[\text{If } t \in (\mu_{j_{k+1}}, \mu_{j_k}), \text{ then } \mu_{j_1} \geq \dots \geq \mu_{j_k} > t > \mu_{j_{k+1}} \geq \dots \geq \mu_{j_n} \right]$$

Extensions to the space of product measures

- **Set-function:** $\mathcal{X}_i = \{0, 1\}$
 - $[0, 1] \approx$ set of probability distributions on $\{0, 1\}$: $\mu_i = \mathbb{P}(X_i = 1)$
 - Lovász extension: for $\mu \in [0, 1]^n$ such that $\mu_{j_1} \geq \dots \geq \mu_{j_n}$

$$\begin{aligned} h(\mu) &= \sum_{k=1}^n \mu_{j_k} [H(e_{j_1} + \dots + e_{j_k}) - H(e_{j_1} + \dots + e_{j_{k-1}})] \\ &= (1 - \mu_{j_1})H(0) + \sum_{k=1}^{n-1} (\mu_{j_k} - \mu_{j_{k+1}})H(e_{j_1} + \dots + e_{j_k}) + \mu_{j_n}H(1_n) \\ &= \mathbb{E}[H(1_{\mu_1 \geq t}, \dots, 1_{\mu_n \geq t})] \text{ for } t \text{ uniform in } [0, 1] \end{aligned}$$

- Lovász extension = relaxation on **product measures**
 - Continuous variable $\mu = (\mu_1, \dots, \mu_n) \in \prod_{i=1}^n [0, 1]$
 - $t \mapsto 1_{\mu_i \geq t}$ is the **inverse cumulative distribution function** of μ_i

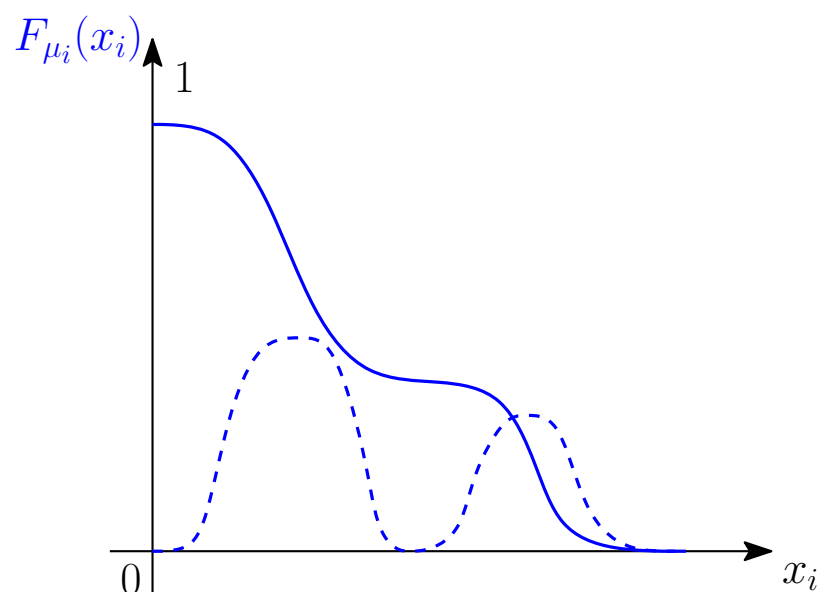
Extensions to the space of product measures

View 1: thresholding cumulative distrib. functions

- Given a probability distribution $\mu_i \in \mathcal{P}(\mathcal{X}_i)$
 - (reversed) cumulative distribution function $F_{\mu_i} : \mathcal{X}_i \rightarrow [0, 1]$ as

$$F_{\mu_i}(x_i) = \mu_i(\{y_i \in \mathcal{X}_i, y_i \geq x_i\}) = \mu_i([x_i, +\infty)) \in [0, 1]$$

- and its “inverse”: $F_{\mu_i}^{-1}(t) = \sup\{x_i \in \mathcal{X}_i, F_{\mu_i}(x_i) \geq t\} \in \mathcal{X}_i$



Extensions to the space of product measures

View 1: thresholding cumulative distrib. functions

- Given a probability distribution $\mu_i \in \mathcal{P}(\mathcal{X}_i)$
 - (reversed) cumulative distribution function $F_{\mu_i} : \mathcal{X}_i \rightarrow [0, 1]$ as

$$F_{\mu_i}(x_i) = \mu_i(\{y_i \in \mathcal{X}_i, y_i \geq x_i\}) = \mu_i([x_i, +\infty)) \in [0, 1]$$

- and its “inverse”: $F_{\mu_i}^{-1}(t) = \sup\{x_i \in \mathcal{X}_i, F_{\mu_i}(x_i) \geq t\} \in \mathcal{X}_i$

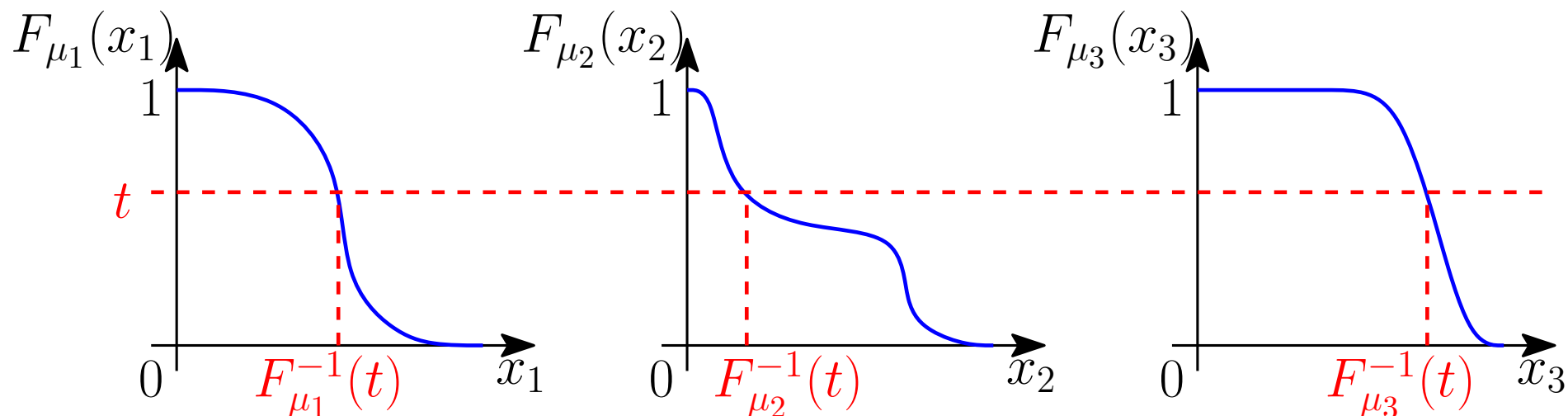
- **“Continuous” extension**

$$\forall \mu \in \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i), \quad h(\mu_1, \dots, \mu_n) = \int_0^1 H[F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)] dt$$

- For finite sets, can be computed by sorting *all* values of $F_{\mu_i}(x_i)$
- Equal to the Lovász extension for set-functions

Extensions to the space of product measures

View 1: thresholding cumulative distrib. functions



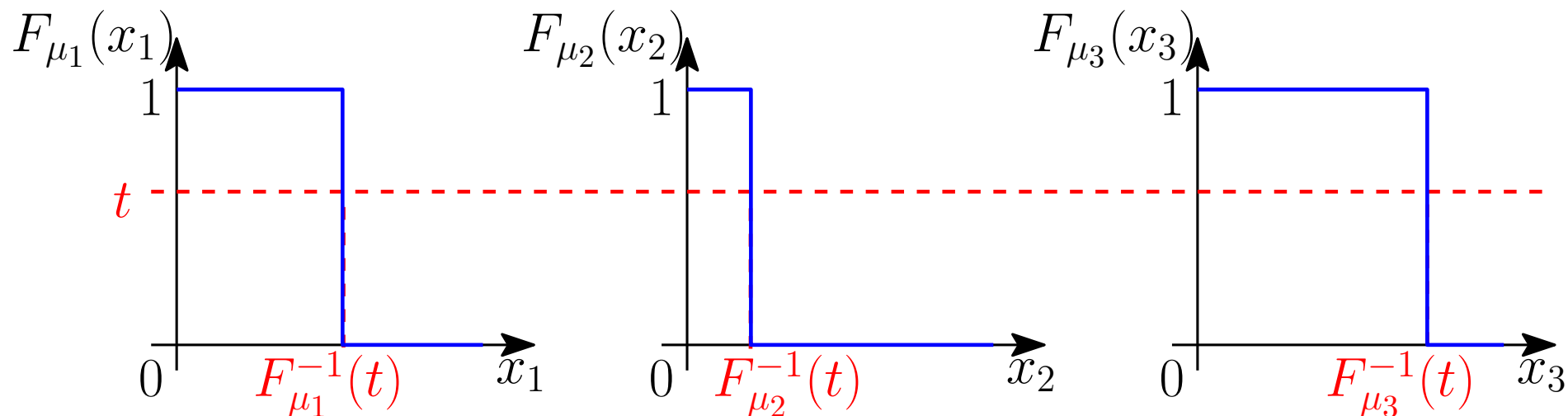
- “Continuous” extension

$$\forall \mu \in \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i), \quad h(\mu_1, \dots, \mu_n) = \int_0^1 H[F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)] dt$$

- For finite sets, can be computed by sorting *all* values of $F_{\mu_i}(x_i)$
- Equal to the Lovász extension for set-functions

Extensions to the space of product measures

View 1: thresholding cumulative distrib. functions



- “Continuous” extension

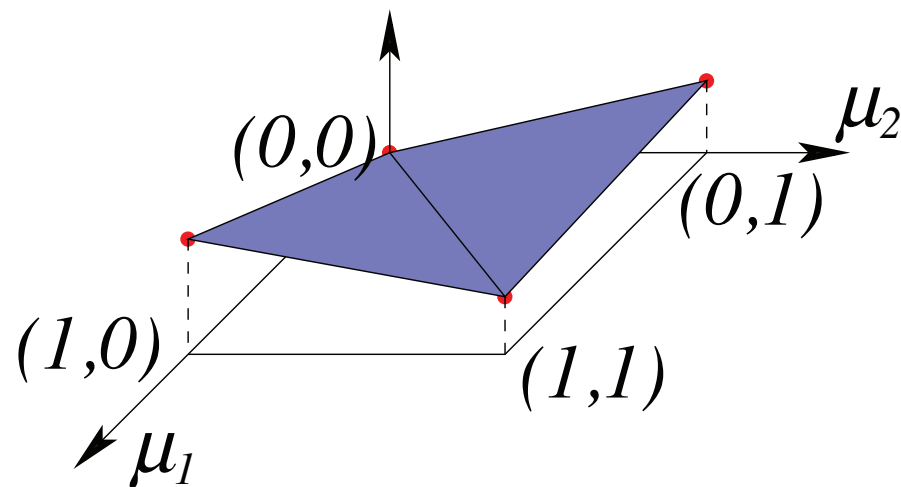
$$\forall \mu \in \prod_{i=1}^n \mathcal{P}(X_i), \quad h(\mu_1, \dots, \mu_n) = \int_0^1 H[F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)] dt$$

- For finite sets, can be computed by sorting *all* values of $F_{\mu_i}(x_i)$
- Equal to $H(x_1, \dots, x_n)$ when $\mu_i = \delta_{x_i}$ for all i

Extensions to the space of product measures

View 2: convex closure

- Given **any** function H on $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$
 - Known value $H(x)$ for any “extreme points” of product measures (i.e., all Diracs δ_x at any $x \in \mathcal{X}$)
 - Convex closure $\tilde{h} =$ largest convex lower bound
 - Minimizing H and its convex closure \tilde{h} is equivalent



Extensions to the space of product measures

View 2: convex closure

- Given

Computation of the convex envelope

- Need to compute the bi-conjugate of

$a : \mu \mapsto H(x)$ if $\mu = \delta_x$ for some $x \in \mathcal{X}$, and $+\infty$ otherwise

- Step 1: compute $a^*(w) = \sup_{\mu} \langle \mu, w \rangle - a(\mu)$ for $w \in \prod_{i=1}^n \mathbb{R}^{\mathcal{X}_i}$

$$\begin{aligned} a^*(w) &= \sup_{x \in \mathcal{X}} \sum_{i=1}^n w_i(x_i) - H(x) = \sup_{\gamma \in \mathcal{P}(\mathcal{X})} \sum_{x \in \mathcal{X}} \gamma(x) \left\{ \sum_{i=1}^n w_i(x_i) - H(x) \right\} \\ &= \sup_{\gamma \in \mathcal{P}(\mathcal{X})} \left\{ \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} w_i(x_i) \gamma_i(x_i) - \sum_{x \in \mathcal{X}} \gamma(x) H(x) \right\} \end{aligned}$$

– with $\gamma_i(x_i) = \sum_{x_j, j \neq i} \gamma(x_1, \dots, x_n)$ the i -th marginal of γ

Computation of the convex envelope

- Step 1: $a^*(w) = \sup_{\gamma \in \mathcal{P}(\mathcal{X})} \left\{ \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} w_i(x_i) \gamma_i(x_i) - \sum_{x \in \mathcal{X}} \gamma(x) H(x) \right\}$

- Step 2: compute $a^{**}(\mu) = \sup_w \langle w, \mu \rangle - a^*(w)$ for $\mu \in \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)$

$$a^{**}(\mu) = \sup_w \langle w, \mu \rangle - \sup_{\gamma \in \mathcal{P}(\mathcal{X})} \left\{ \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} w_i(x_i) \gamma_i(x_i) - \sum_{x \in \mathcal{X}} \gamma(x) H(x) \right\}$$

$$= \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \sup_w \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} w_i(x_i) (\mu_i(x_i) - \gamma_i(x_i)) + \sum_{x \in \mathcal{X}} \gamma(x) H(x)$$

- Thus $a^{**}(\mu) = \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} H(x) d\gamma(x)$ such that $\forall i, \gamma_i(x_i) = \mu_i(x_i)$

Extensions to the space of product measures

View 2: convex closure

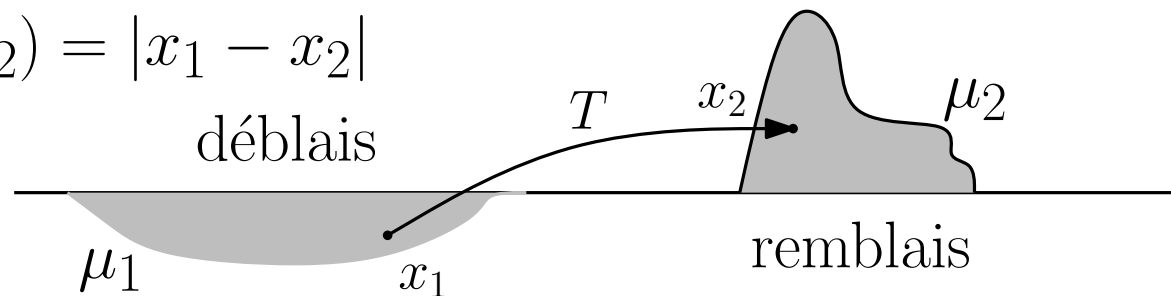
- Given **any** function H on $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$
 - Known value $H(x)$ for any “extreme points” of product measures (i.e., all Diracs δ_x at any $x \in \mathcal{X}$)
 - Convex closure $\tilde{h} =$ largest convex lower bound
 - Minimizing H and its convex closure \tilde{h} is equivalent
- “Closed-form” formulation: $\tilde{h}(\mu_1, \dots, \mu_n) = \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} H(x) d\gamma(x),$
 - with respect to all prob. measures γ on \mathcal{X} such that $\gamma_i(x_i) = \mu_i(x_i)$
 - **Multi-marginal optimal transport**

Optimal transport: from Monge to Kantorovich

- **Monge formulation** (“La théorie des déblais et des remblais”, 1781)

- Transforming a measure μ_1 to μ_2 that (a) preserves local mass and (b) minimize transportation cost $\int_{x_1} c(x_1, T(x_1)) d\mu_1(x_1)$

$$c(x_1, x_2) = |x_1 - x_2|$$



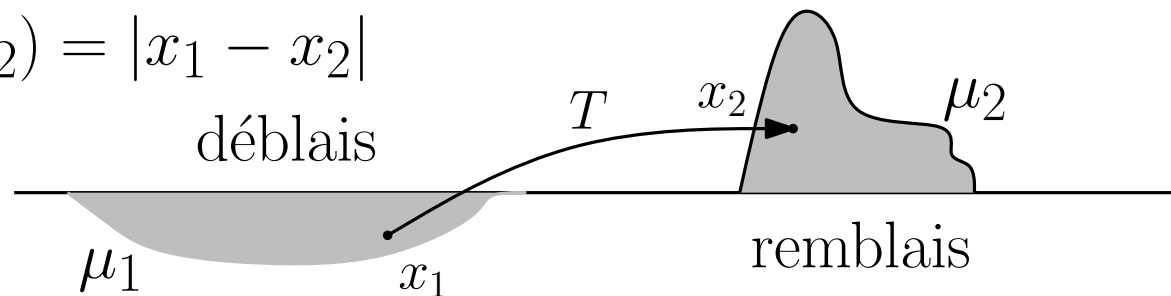
- Optimal transport map T may not always exist
- Discrete case: earth's mover distance

Optimal transport: from Monge to Kantorovich

- **Monge formulation** (“La théorie des déblais et des remblais”, 1781)

- Transforming a measure μ_1 to μ_2 that (a) preserves local mass and (b) minimize transportation cost $\int_{x_1} c(x_1, T(x_1)) d\mu_1(x_1)$

$$c(x_1, x_2) = |x_1 - x_2|$$



- Optimal transport map T may not always exist
- Discrete case: earth's mover distance

- **Kantorovich formulation (1942)**

- **Convex relaxation** on space of probability measures

Optimal transport: from two to multiple marginals

- **Kantorovich formulation (1942)**

- **Convex relaxation** on space of probability measures $\gamma \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$
- Prescribed marginals $\gamma_1 = \mu_1$ and $\gamma_2 = \mu_2$
- Minimum cost $\int_{\mathcal{X}_1 \times \mathcal{X}_2} c(x_1, x_2) d\gamma(x_1, x_2)$

- **Properties**

- Monge formulation with distribution of $(x_1, T(x_1))$
- Wasserstein distance between measures with $c(x_1, x_2) = |x_1 - x_2|^p$
- Relationship with copulas
- See Villani (2008); Santambrogio (2015)

Optimal transport: from two to multiple marginals

- **Kantorovich formulation (1942)**

- **Convex relaxation** on space of probability measures $\gamma \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$
- Prescribed marginals $\gamma_1 = \mu_1$ and $\gamma_2 = \mu_2$
- Minimum cost $\int_{\mathcal{X}_1 \times \mathcal{X}_2} c(x_1, x_2) d\gamma(x_1, x_2)$

- **Properties**

- Monge formulation with distribution of $(x_1, T(x_1))$
- Wasserstein distance between measures with $c(x_1, x_2) = |x_1 - x_2|^p$
- Relationship with copulas
- See Villani (2008); Santambrogio (2015)

- **Extension to multiple marginals**

- Minimize $\int_{\mathcal{X}} H(x) d\gamma(x)$ with respect to all prob. measures γ on \mathcal{X} such that $\gamma_i(x_i) = \mu_i(x_i)$ for all $i \in \{1, \dots, n\}$

Extensions to the space of product measures

Combining the two views

- **View 1: thresholding cumulative distribution functions**

- + closed form computation for any H , always an extension
- not convex

- **View 2: convex closure**

- + convex for any H , allows minimization of H
- not computable, may not be an extension

Extensions to the space of product measures

Combining the two views

- **View 1: thresholding cumulative distribution functions**

- + closed form computation for any H , always an extension
- not convex

- **View 2: convex closure**

- + convex for any H , allows minimization of H
- not computable, may not be an extension

- **Submodularity**

- The two views are equivalent
- Direct proof through optimal transport
- All results from submodular set-functions go through

Kantorovich optimal transport in one dimension

- **Theorem** (Carlier, 2003): If H is submodular, then

$$\inf_{\gamma \in \mathcal{P}(X)} \int_X H(x) d\gamma(x) \text{ such that } \forall i, \gamma_i = \mu_i$$

is equal to $\int_0^1 H[F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)] dt$

Kantorovich optimal transport in one dimension

- **Theorem** (Carlier, 2003): If H is submodular, then

$$\inf_{\gamma \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} H(x) d\gamma(x) \text{ such that } \forall i, \gamma_i = \mu_i$$

is equal to $\int_0^1 H[F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)] dt$

- **Proof/intuition for $n = 2$ for the Monge problem**

(a) Assume for simplicity atomless measures

(b) The following increasing map is natural $F_{\mu_2}^{-1} \circ F_{\mu_1} : \mathcal{X}_1 \rightarrow \mathcal{X}_2$

(c) This is the only increasing map

(d) **Transport maps always increasing when H submodular**

– If $x_1 < x'_1$ mapped to $x_2 > x'_2$, then exchanging x_2 and x'_2 would increase cost by $H(x_1, x'_2) + H(x'_1, x_2) - H(x_1, x_2) - H(x'_1, x'_2) \leq 0$

Duality - Subgradients of extension

- **General duality**

$$h(\mu) = \sup_w \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} w_i(x_i) \mu_i(x_i) - \sup_{x \in \mathcal{X}} \left\{ \sum_{i=1}^n w_i(x_i) - H(x) \right\}$$

- **Subgradients from “greedy algorithm”**

- Sort all values of $F_{\mu_i}(x_i)$ for $i \in \{1, \dots, n\}$ and $x_i \in \mathcal{X}_i$
- Get a subgradient w by taking differences of values of H
- See Bach (2015) for more details

- **Extensions of various submodular polytopes**

Submodular functions

Links with convexity (Bach, 2015)

1. H is submodular if and only if h is convex

2. If H is submodular, then

$$\min_{x \in \prod_{i=1}^n \mathcal{X}_i} H(x) = \min_{\mu \in \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)} h(\mu)$$

3. If H is submodular, then a **subgradient** of h at any μ may be computed by a “greedy algorithm”

Submodular functions

Links with convexity (Bach, 2015)

1. H is submodular if and only if h is convex

2. If H is submodular, then

$$\min_{x \in \prod_{i=1}^n \mathcal{X}_i} H(x) = \min_{\mu \in \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)} h(\mu)$$

3. If H is submodular, then a **subgradient** of h at any μ may be computed by a “greedy algorithm”

- Submodular functions may be minimized in polynomial time with similar algorithms than for the binary case
- NB: existing reduction to submodular set-functions defined on a ring family (Schrijver, 2000)

Outline

1. Submodular set-functions

- Definitions, examples
- Links with convexity through Lovász extension
- Minimization by convex optimization

2. From discrete to continuous domains

- Nonpositive second-order derivatives
- Invariances and examples
- Extensions on product measures through optimal transport

3. Minimization of continuous submodular functions

- Subgradient descent
- Frank-Wolfe optimization

Minimization of submodular functions

Projected subgradient descent

- **For simplicity:** discretizing all sets \mathcal{X}_i , $i = 1, \dots, n$ to k elements
- Assume **Lipschitz-continuity:** $\forall x, e_i, |H(x + e_i) - H(x)| \leq B$
 - Fact: subgradients of h bounded by B in ℓ_∞ -norm
- **Projected subgradient descent**
 - Convergence rate of $O(nkB/\sqrt{t})$ after t iterations
 - Cost of each iteration $O(nk \log(nk))$
 - Reasonable scaling with respect to discretization

$$\tilde{O}\left(\frac{n^3}{\varepsilon^3}\right) \text{ for continuous domains}$$

Minimization of submodular functions

Frank-Wolfe / conditional gradient

- **Submodular set-functions:** $\mathcal{X}_i = \{0, 1\}$
 - (C) : $\min_{\mu \in [0,1]^n} h(\mu)$ non-smooth convex
 - Solve instead (S) : $\min_{\mu \in \mathbb{R}^n} h(\mu) + \frac{1}{2}\|\mu\|^2$ (strongly convex)
 - Fact: level sets of (S) obtained from minimizers of $H(x) + \lambda x^\top \mathbf{1}_n$

Minimization of submodular functions

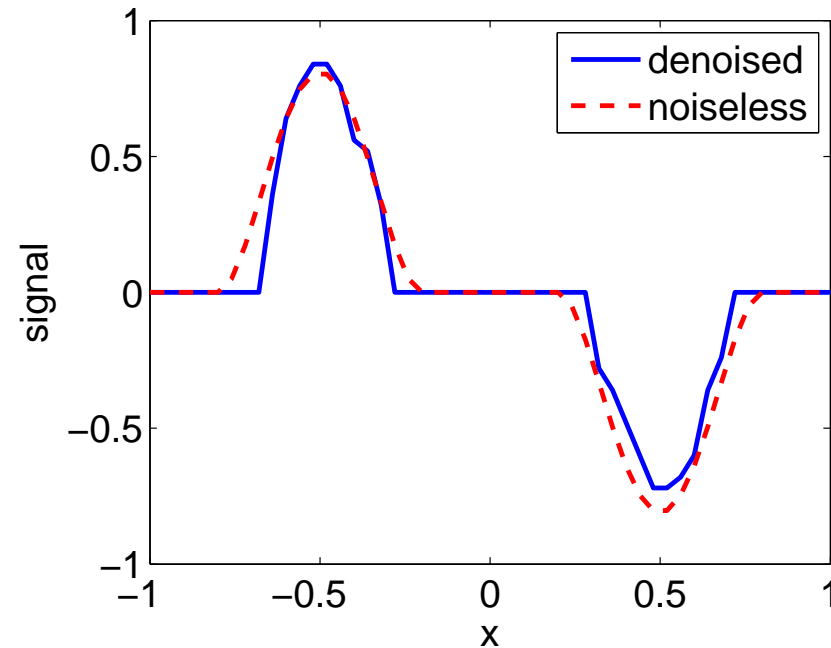
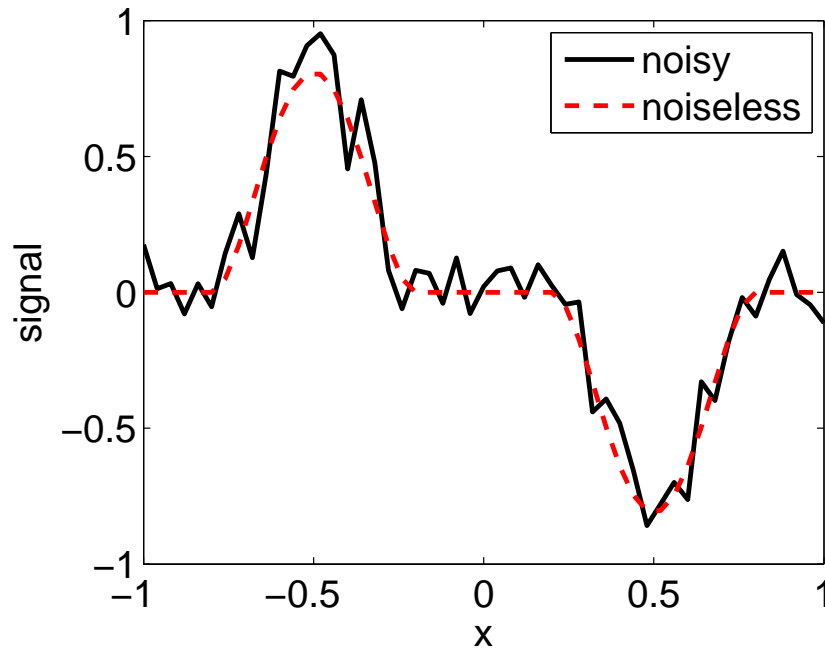
Frank-Wolfe / conditional gradient

- **Submodular set-functions:** $\mathcal{X}_i = \{0, 1\}$
 - (C) : $\min_{\mu \in [0,1]^n} h(\mu)$ non-smooth convex
 - Solve instead (S) : $\min_{\mu \in \mathbb{R}^n} h(\mu) + \frac{1}{2}\|\mu\|^2$ (strongly convex)
 - Fact: level sets of (S) obtained from minimizers of $H(x) + \lambda x^\top \mathbf{1}_n$
- **Extension to all submodular functions**
 - (C) : $\min_{\mu \in \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)} h(\mu)$
 - Solve instead (S) : $\min_{\mu \in \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)} h(\mu) + \sum_{i=1}^n \varphi_i(\mu_i)$
 - $\varphi(\mu_i)$ defined through optimal transport with a submodular cost $c_i(x_i, t)$ between μ_i and the uniform distribution on $[0, 1]$
 - $\varphi(\mu_i)$ can be strongly convex
 - Level sets of (S) obtained from minimizers of $H(x) + \sum_{i=1}^n c_i(x_i, t)$

Empirical simulations (online code)

- Signal processing example: $H : [-1, 1]^n \rightarrow \mathbb{R}$ with $\alpha < 1$

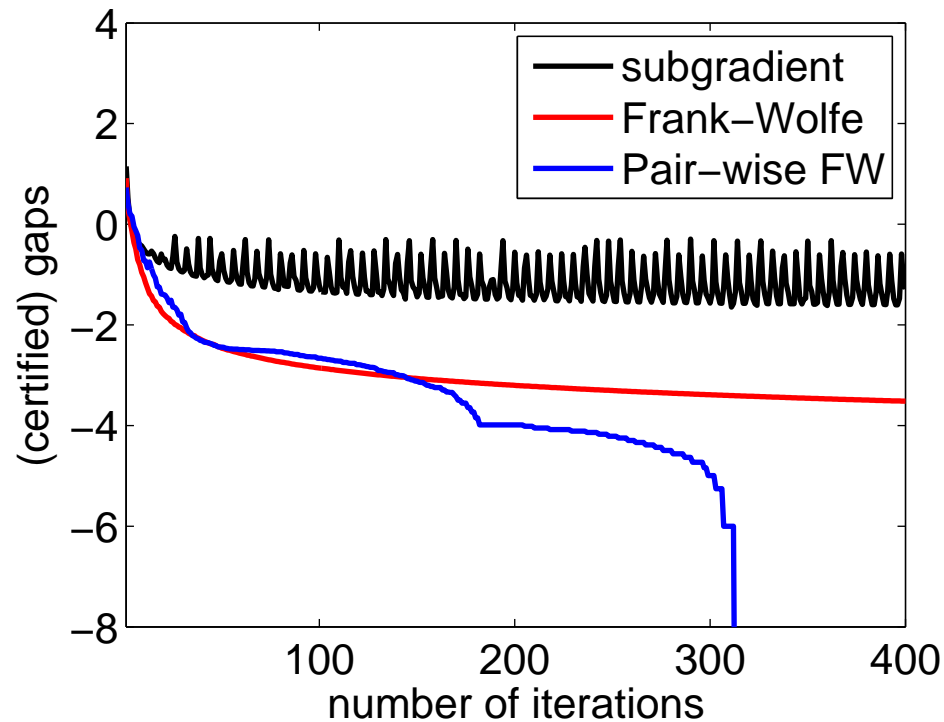
$$H(x) = \frac{1}{2} \sum_{i=1}^n (x_i - z_i)^2 + \lambda \sum_{i=1}^n |x_i|^\alpha + \mu \sum_{i=1}^{n-1} (x_i - x_{i+1})^2$$



Empirical simulations (online code)

- Signal processing example: $H : [-1, 1]^n \rightarrow \mathbb{R}$ with $\alpha < 1$

$$H(x) = \frac{1}{2} \sum_{i=1}^n (x_i - z_i)^2 + \lambda \sum_{i=1}^n |x_i|^\alpha + \mu \sum_{i=1}^{n-1} (x_i - x_{i+1})^2$$

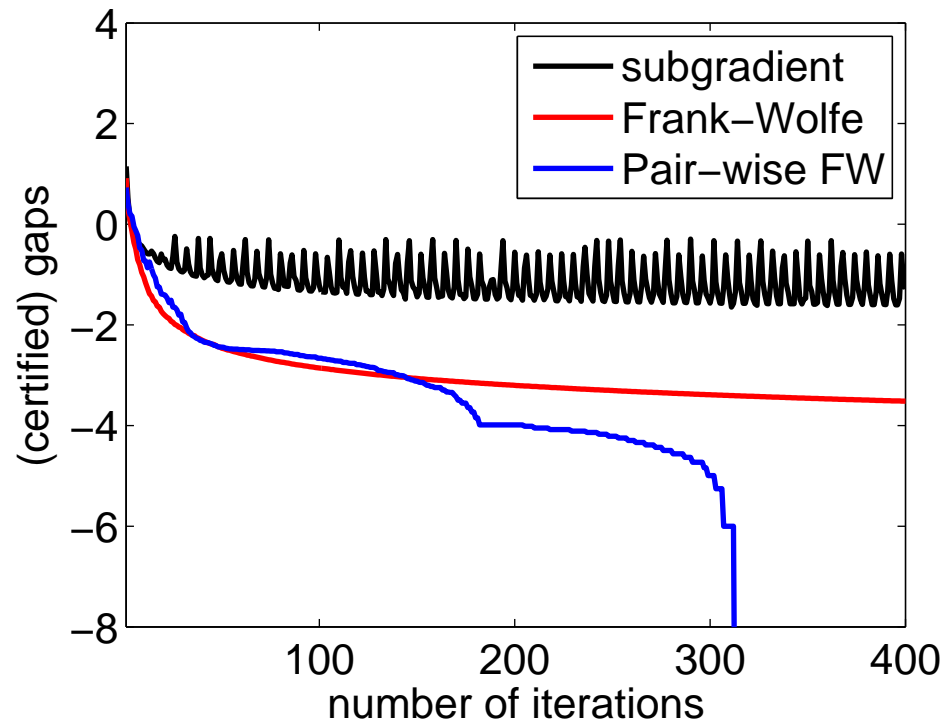


- Pair-wise Frank-Wolfe (Lacoste-Julien and Jaggi, 2015)

Empirical simulations (online code)

- Signal processing example: $H : [-1, 1]^n \rightarrow \mathbb{R}$ with $\alpha < 1$

$$H(x) = \frac{1}{2} \sum_{i=1}^n (x_i - z_i)^2 + \lambda \sum_{i=1}^n |x_i|^\alpha + \mu \sum_{i=1}^{n-1} (x_i - x_{i+1})^2$$



- Pair-wise Frank-Wolfe (Lacoste-Julien and Jaggi, 2015)

Conclusion

- **Submodular function and convex optimization**
 - From discrete to continuous domains
 - Extensions to product measures
 - Direct link with one-dimensional multi-marginal optimal transport

Conclusion

- **Submodular function and convex optimization**

- From discrete to continuous domains
- Extensions to product measures
- Direct link with one-dimensional multi-marginal optimal transport

- **On-going work and extensions**

- Optimal transport beyond submodular functions
- Beyond discretization
- Beyond minimization
- Sums of submodular functions and convex functions
- Sums of simple submodular functions (Jegelka et al., 2013)
- Mean-field inference in log-supermodular models (Djolonga and Krause, 2015)

References

- F. Bach. Learning with Submodular Functions: A Convex Optimization Perspective. Technical Report 00645271, HAL, 2013.
- F. Bach. Submodular functions: from discrete to continuous domains. Technical Report 1511.00394-v2, HAL, 2015.
- G. Carlier. On a class of multidimensional optimal transportation problems. *Journal of Convex Analysis*, 10(2):517–530, 2003.
- A. Chambolle. Total variation minimization and a class of binary MRF models. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 136–152. Springer, 2005.
- G. Choquet. Theory of capacities. *Ann. Inst. Fourier*, 5:131–295, 1954.
- J. Djolonga and A. Krause. Scalable variational inference in log-supermodular models. In *International Conference on Machine Learning (ICML)*, 2015.
- S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.
- M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.
- S. Iwata, L. Fleischer, and S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.
- S. Jegelka, F. Bach, and S. Sra. Reflection methods for user-friendly submodular optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

- Sunyoung Kim and Masakazu Kojima. Exact solutions of some nonconvex quadratic optimization problems via sdp and socp relaxations. *Computational Optimization and Applications*, 26(2): 143–154, 2003.
- A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proc. UAI*, 2005.
- S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- G. G. Lorentz. An inequality for rearrangements. *American Mathematical Monthly*, 60(3):176–179, 1953.
- L. Lovász. Submodular functions and convexity. *Mathematical programming: The state of the art, Bonn*, pages 235–257, 1982.
- P. Milgrom and C. Shannon. Monotone comparative statics. *Econometrica: Journal of the Econometric Society*, pages 157–180, 1994.
- J.B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.
- F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer, 2015.
- A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.
- M. Seeger. On the submodularity of linear experimental design, 2009. http://lapma1.epfl.ch/papers/subm_lindesign.pdf.
- D. M. Topkis. Minimizing a submodular function on a lattice. *Operations research*, 26(2):305–321,

1978.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

A. Weller. Bethe and related pairwise entropy approximations. In *Uncertainty in Artificial Intelligence (UAI)*, 2015.