Efficient Algorithms for Non-convex Isotonic Regression through Submodular Optimization

Francis Bach
INRIA, École Normale Supérieure, PSL, Paris

Isotonic Regression

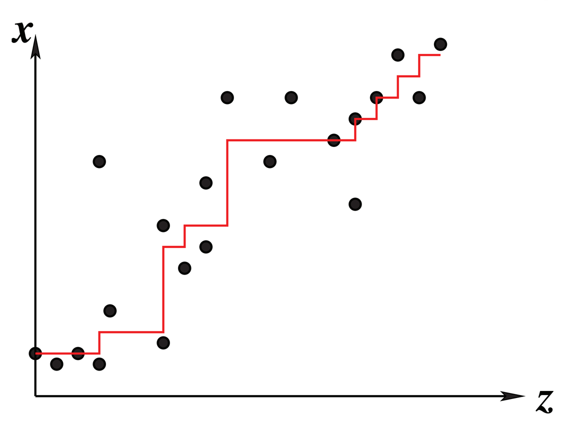
- Imposing ordering constraints on estimation problems
- Signal processing, machine learning and statistics
  - Probability calibration for classification
  - Interpretability

Constrained optimization problem

$$\min_{x\in\mathbb{R}^n} H(x) \text{ such that } \forall (i,j)\in E,\ x_i \geqslant x_j,$$

where $E \subset \{1,\dots,n\}^2$ is a directed acyclic graph.

- Most classical example: separable quadratic cost



$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-z\|_2^2$$
such that $x_1 \leqslant \cdots \leqslant x_n$

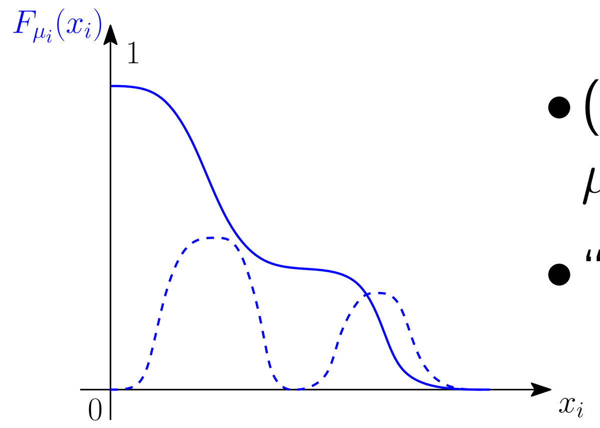
- Convex function $H \Rightarrow$ convex optimization problem
  - Efficient first-order algorithms
- Non-convex function H
  - No efficient algorithms general

Submodular Functions

- Definition ([1]): $\forall x\in\mathbb{R}^n,\ \forall i\neq j,\ \dfrac{\partial^2 H}{\partial x_i \partial x_j}(x) \leqslant 0$
- Example 1 (Laplacian): $H(x) = \sum_{i,j=1}^n w_{ij}(x_i - x_j)^2$
- Example 2 (Separable functions): $H(x) = \sum_{i=1}^n h_i(x_i)$
  - For any (e.g., non-convex) functions $h_i : \mathbb{R}\to\mathbb{R}$

Convex Extension with Measures

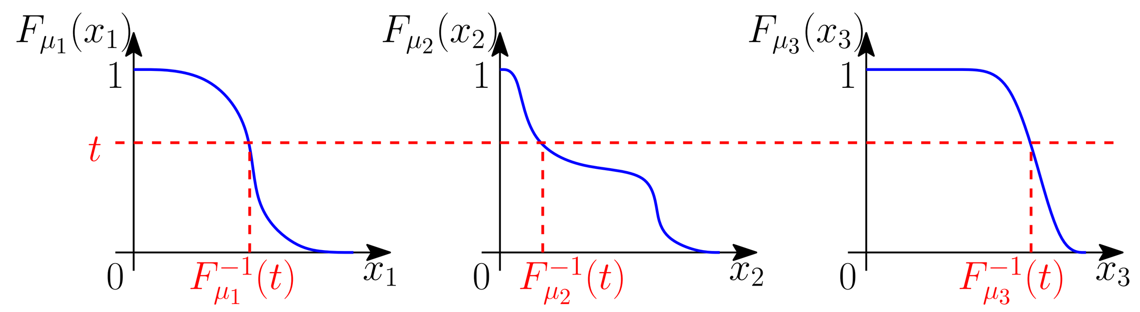
- $\mathcal{P}(\mathbb{R})$ = set of of Radon probability measures on $\mathbb{R}$



- (reverse) cumulative distribution function for $\mu_i \in \mathcal{P}(\mathbb{R})$: $F_{\mu_i}(x_i) = \mu_i([x_i,+\infty))$
- "Inverse" CDF $F_{\mu_i}^{-1}$

- Extension to $\mu = (\mu_1,\dots,\mu_n)$–see [2]:

$$h(\mu_1,\dots,\mu_m) = \int_0^1 H\big[F_{\mu_1}^{-1}(t),\dots,F_{\mu_n}^{-1}(t)\big]dt$$

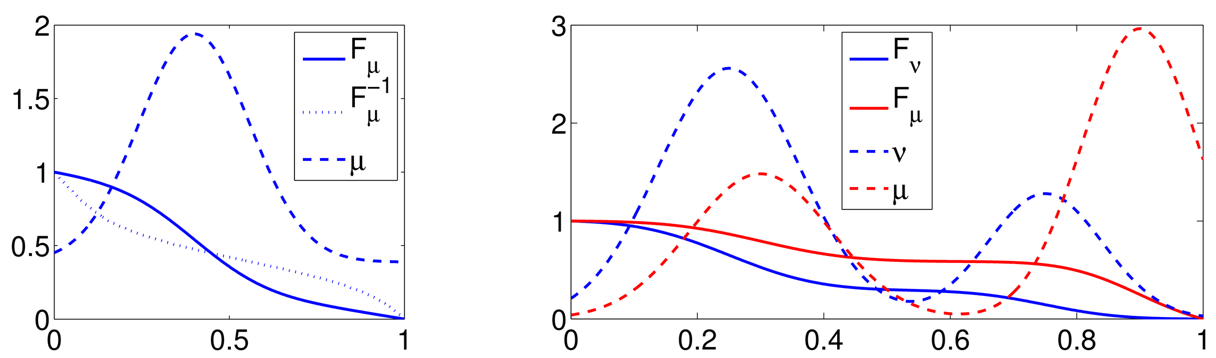


- Property 1: H submodular $\Leftrightarrow$ h convex
- Property 2: minimizers of H can be recovered from minimizers of h
- Consequence: Submodular function minimization is a convex optimization problem

Extension for Isotonic Constraints

- First-order stochastic dominance between $\mu,\nu\in\mathcal{P}(\mathbb{R})$:

$$\mu \succcurlyeq \nu \text{ if and only if } \forall x\in\mathbb{R},\ F_\mu(x) \geqslant F_\nu(x)$$



- Equivalent optimization problem

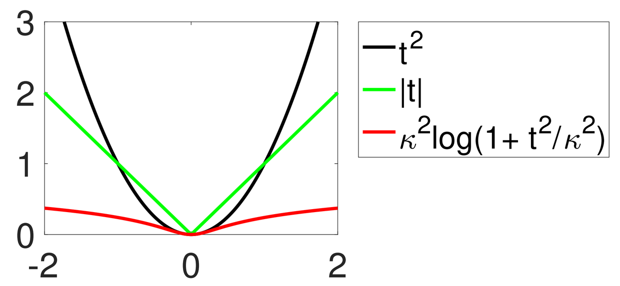
$$\min_{\mu\in\mathcal{P}(\mathbb{R})^n} h(\mu) \text{ such that } \forall (i,j)\in E,\ \mu_i \succcurlyeq \mu_j$$

Discretization Algorithms

- Discretize each interval with k values
  - Approximation $O(n/k)$
  - Number of queries of H is $O(tnk)$ for t operations of projected subgradient descent
  - Projection based on (efficient) quadratic cost isotonic regression [3]
  - Extra approximation factor $O(n/\sqrt{t})$
- With $k = n/\varepsilon$ and $t = n^2/\varepsilon^2$
  - Error ε and number of queries $O(n^4/\varepsilon^3)$
- For improved behavior, see paper

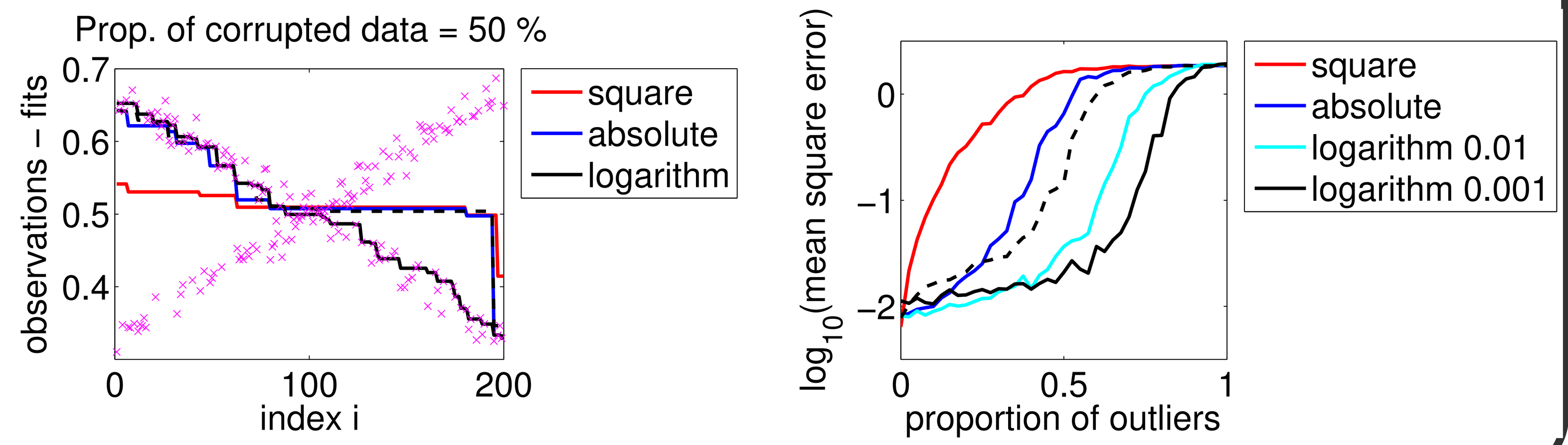
Experiments

- Robust isotonic regression: $H(x) = \frac{1}{n}\sum_{i=1}^n G(x_i - z_i)$



- $G(t) = |t|$ or $|t|^2 \Rightarrow$ convex problem
- $G(t) = \kappa^2\log(1+\frac{t^2}{\kappa^2}) \Rightarrow$ submodular problem
- Vanishing gradients $\Rightarrow$ Robustness

- Data with adversarial perturbations
  - Goal: learn decreasing functions
  - Perturbation by noise and increasing function (observations in pink)



References

[1] D. M. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26(2):305– 321, 1978.

[2] F. Bach. Submodular functions: from discrete to continuous domains. *Mathematical Programming*, 2018

[3] Q. F. Stout. Isotonic regression via partitioning. *Algorithmica*, 66(1):93–112, 2013.