Frank-Wolfe, Shapley-Folkman

& Nonconvex Separable Problems

Alexandre d'Aspremont,

CNRS & D.I., École Normale Supérieure.

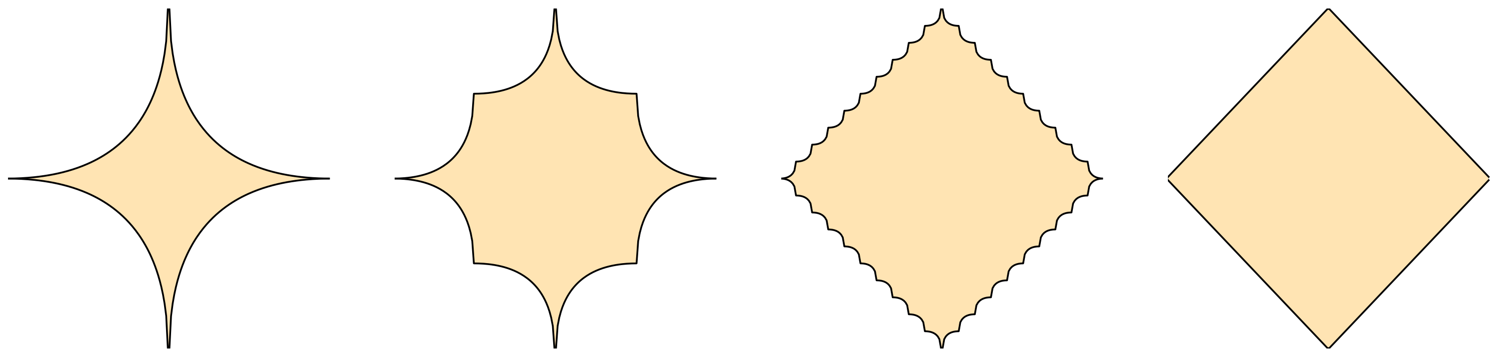
With Armin Askari (Voleon), Benjamin Dubois-Taine (INRIA),
Laurent El Ghaoui (UC Berkeley) and Quentin Rebjock (EPFL).

Jobs

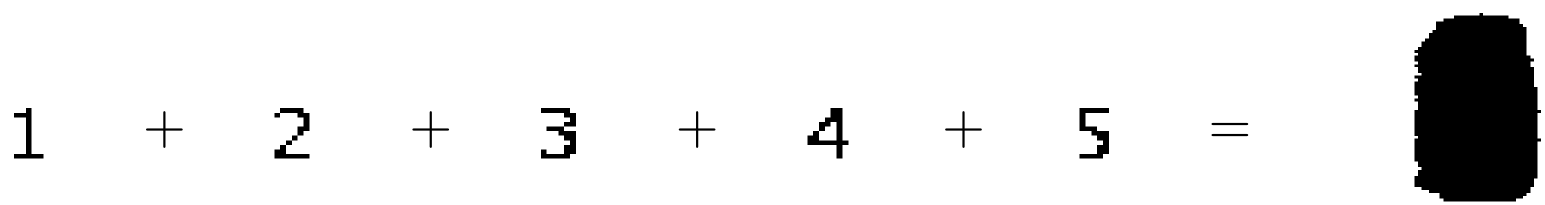
Postdoc positions in **ML / Optimization**.

At CNRS, INRIA, Ecole Normale Supérieure in Paris.

Shapley-Folkman Theorem



The $\ell_{1/2}$ ball, Minkowsi average of two and ten balls, convex hull.



Minkowsi sum of five first digits (obtained by sampling).

Shapley-Folkman Theorem

Shapley-Folkman Theorem [Starr, 1969, Emerson and Greenleaf, 1969]

Suppose $V_i \subset \mathbb{R}^d$, $i = 1, \ldots, n$, *and*

$$x \in \sum_{i=1}^{n} \mathbf{Co}(V_i)$$

then

$$x \in \sum_{[1,n]\setminus\mathcal{S}} V_i + \sum_{\mathcal{S}} \mathbf{Co}(V_i)$$

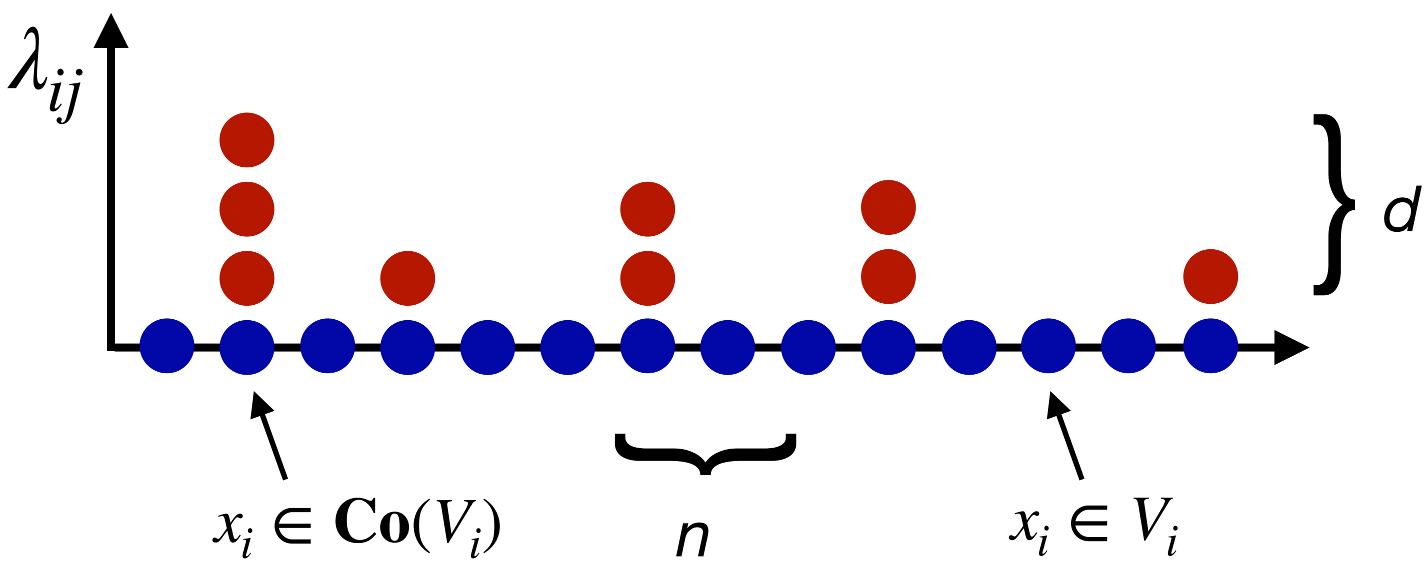
for some $|\mathcal{S}| \leq d$.

Shapley-Folkman Theorem: Carathéodory

Proof sketch. Write $x \in \sum_{i=1}^{n} \mathbf{Co}(V_i)$, or

$$\begin{pmatrix} x \\ \mathbf{1}_n \end{pmatrix} = \sum_{i=1}^{n} \sum_{j=1}^{d+1} \lambda_{ij} \begin{pmatrix} v_{ij} \\ e_i \end{pmatrix}, \quad \text{for } \lambda \geq 0,$$

Conic Carathéodory then yields representation with at most $n + d$ nonzero coefficients. Use a pigeonhole argument



Number of nonzero λ_{ij} controls gap with convex hull.

Shapley-Folkman: a Quantization Result

One Line Proof. Suppose $x \in \frac{\sum_{i=1}^n V}{n}$ for some $V \in \mathbb{R}^d$.

- Write x as
$$x = \sum_{i=1}^{T} \frac{\mu_i}{n}\, v_i, \quad \text{where } \sum_{i=1}^{T} \frac{\mu_i}{n} = 1$$
for $v_i \in V$, where $\mu_i \geq 0$ is the number of times v_i is repeated in $\sum_{i=1}^n V$.

- As $n \to \infty$, we can approximate any point in $\mathbf{Co}(V)$
$$\tilde{x} = \sum_{i=1}^{d+1} \lambda_i v_i, \quad \text{where } \sum_{i=1}^{d+1} \lambda_i = 1,\ \lambda \geq 0,$$
by a point of $\frac{\sum_{i=1}^n V}{n}$, with arbitrarily small **quantization error $1/n$.**

Shapley-Folkman

Consequences.

- If the sets $V_i \subset \mathbb{R}^d$ are uniformly bounded with $\mathrm{rad}(V_i) \leq R$, then

$$d_H\left(\frac{\sum_{i=1}^n V_i}{n}, \mathbf{Co}\left(\frac{\sum_{i=1}^n V_i}{n}\right)\right) \leq R\frac{\sqrt{\min\{n,d\}}}{n}$$

  where $\mathrm{rad}(V) = \inf_{x\in V}\sup_{y\in V}\|x-y\|$.

- Holds for many other nonconvexity measures (e.g. volume deficit) [Fradelizi et al., 2017].

Outline

- The Shapley-Folkman Theorem
- **Duality Gap Bounds**
- Primalization
- Applications

Nonconvex Optimization

Separable nonconvex problem. Solve

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n f_i(x_i) \\ \text{subject to} & Ax \leq b, \end{array} \tag{P}$$

in the variables $x_i \in \mathbb{R}^{d_i}$ with $d = \sum_{i=1}^n d_i$, where f_i are lower semicontinuous and $A \in \mathbb{R}^{m \times d}$.

Take the dual twice to form a **convex relaxation**,

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n f_i^{**}(x_i) \\ \text{subject to} & Ax \leq b \end{array} \tag{CoP}$$

in the variables $x_i \in \mathbb{R}^{d_i}$.

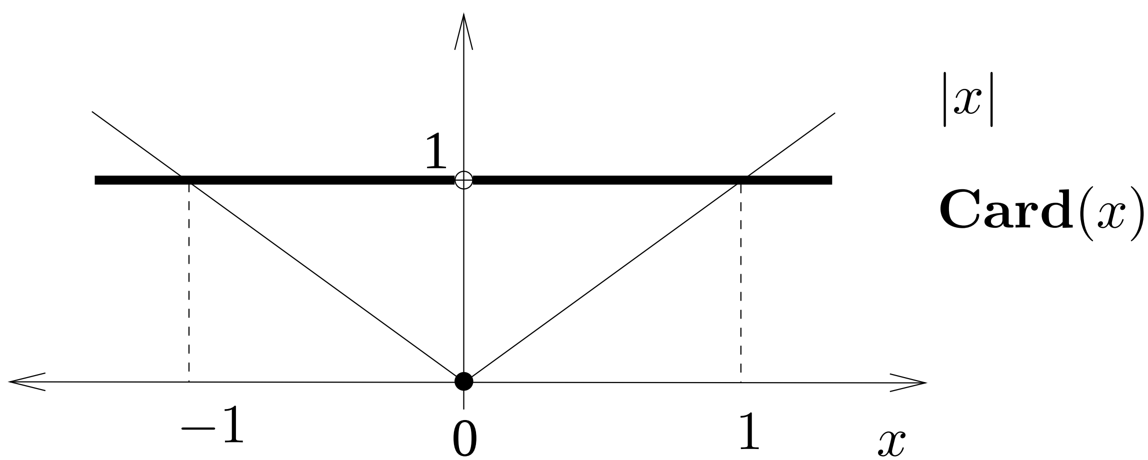
Nonconvex Optimization

Convex envelope. Biconjugate f^{**} satisfies $\mathbf{epi}(f^{**}) = \overline{\mathbf{Co}(\mathbf{epi}(f))}$, which means that

$f^{**}(x)$ and $f(x)$ match at extreme points of $\mathbf{epi}(f^{**})$.

Define **lack of convexity** as $\rho(f) \triangleq \sup_{x \in \mathbf{dom}(f)} \{f(x) - f^{**}(x)\}$.

Example.



The l_1 norm is the convex envelope of $\mathbf{Card}(x)$ in $[-1, 1]$.

Nonconvex Optimization

Writing the **epigraph** of problem (P) as in [Lemaréchal and Renaud, 2001],

$$\mathcal{G}_r \triangleq \left\{ (r_0, r) \in \mathbb{R}^{1+m} : \; \sum_{i=1}^n f_i(x_i) \leq r_0,\; Ax - b \leq r, x \in \mathbb{R}^d \right\},$$

we can write the dual function of (P) as

$$\Psi(\lambda) \triangleq \inf \left\{ r_0 + \lambda^\top r : \; (r_0, r) \in \mathcal{G}_r^{**} \right\},$$

in the variable $\lambda \in \mathbb{R}^m$, where $\mathcal{G}^{**} = \overline{\mathbf{Co}(\mathcal{G})}$ is the **closed convex hull of the epigraph** $\mathcal{G}$**.**

If $\mathcal{G}^{**} = \mathcal{G}$, no duality gap in (P).

Nonconvex Optimization

Epigraph & duality gap. Define

$$\mathcal{F}_i = \left\{ (f_i(x_i), A_i x_i) : \; x_i \in \mathbb{R}^{d_i} \right\} + \mathbb{R}^{m+1}_+$$

where $A_i \in \mathbb{R}^{m \times d_i}$ is the i^{th} block of A.

- **The epigraph of problem (P) can be written as a Minkowski sum of $\mathcal{F}_i$**

$$\mathcal{G}_r = \sum_{i=1}^{n} \mathcal{F}_i + (0, -b) + \mathbb{R}^{m+1}_+$$

- Shapley-Folkman thus shows $f^{**}(x_i) = f(x_i)$ for **all but at most** $m+1$ **terms in the objective.**

- As $n \to \infty$, with $m/n \to 0$, $\mathcal{G}_r$ gets closer to its convex hull $\mathcal{G}_r^{**}$, and the **duality gap becomes negligible.**

Bound on duality gap

Linear constraints. A priori bound on duality gap of

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n f_i(x_i) \\ \text{subject to} & Ax \leq b, \end{array}$$

where $A \in \mathbb{R}^{m \times d}$.

Proposition [Aubin and Ekeland, 1976, Ekeland and Temam, 1999]

A priori bounds on the duality gap *Suppose the functions f_i in (P) satisfy Assumption (. . .). There is a point $x^\star \in \mathbb{R}^d$ at which the primal optimal value of (CoP) is attained, such that*

$$\underbrace{\sum_{i=1}^n f_i^{**}(x_i^\star)}_{CoP} \leq \underbrace{\sum_{i=1}^n f_i(\hat{x}_i^\star)}_{P} \leq \underbrace{\sum_{i=1}^n f_i^{**}(x_i^\star)}_{CoP} + \underbrace{\sum_{i=1}^{m+1} \rho(f_{[i]})}_{\text{gap}}$$

where $\hat{x}^\star$ is an optimal point of (P) and $\rho(f_{[1]}) \geq \rho(f_{[2]}) \geq \ldots \geq \rho(f_{[n]})$.

Bound on duality gap

General result. Consider the separable nonconvex problem

$$\mathrm{h}_P(u) := \begin{array}{ll} \text{min.} & \sum_{i=1}^n f_i(x_i) \\ \text{s.t.} & \sum_{i=1}^n g_i(x_i) \leq b + u \end{array} \qquad \text{(P)}$$

in the variables $x_i \in \mathbb{R}^{d_i}$, with perturbation parameter $u \in \mathbb{R}^m$.

Proposition [Ekeland and Temam, 1999]

A priori bounds on the duality gap *Suppose the functions* f_i, g_{ji} *in problem* (P) *satisfy assumption (...) for* $i = 1, \ldots, n$, $j = 1, \ldots, m$. *Let*

$$\bar{p}_j = (m+1) \max_i \rho(g_{ji}), \quad \text{for } j = 1, \ldots, m$$

then

$$\mathrm{h}_P(\bar{p})^{**} \leq \mathrm{h}_P(\bar{p}) \leq \mathrm{h}_P(0)^{**} + (m+1) \max_i \rho(f_i).$$

where $\mathrm{h}_P(u)^{**}$ *is the optimal value of the dual to* (P).

Outline

- The Shapley-Folkman Theorem
- Duality Gap Bounds
- **Primalization**
- Applications

Primal Solutions

Primalization.

- We have explicit bounds on the duality gap.
- Only **some of the solutions** of the relaxation satisfy the duality gap bounds.
- **How do we efficiently find good primal solutions?**

Randomized algorithm in [Udell and Boyd, 2016].

- Function domains are assumed convex.
- Requires solving a random problem over explicit optimality constraints.

Can we lift these restrictions?

Primal Solutions

Bidual. The bidual is given by

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n f_i^{**}(x_i) \\ \text{subject to} & Ax \leq b, \\ & x_i \in \mathbf{Co}(X_i), \quad i = 1, \ldots, n. \end{array} \tag{CoP}$$

Let $z^* = (v^*, b)$ with v^* the optimal value of (CoP).

There exists $x^* \in \mathbb{R}^d$ such that $Ax^* \leq b$ and $v^* = \sum_{i=1}^n f_i^{**}(x_i^*)$, i.e.

$$z^* \in \sum_{i=1}^n \left\{ \begin{pmatrix} f_i^{**}(x_i) \\ A_i x_i \end{pmatrix} \mid x_i \in \operatorname{conv} X_i \right\} + \mathbb{R}_+^{m+1}. \tag{1}$$

Define $\mathcal{C} = \sum_{i=1}^n \mathcal{C}_i$, with $\mathcal{C}_i = \left\{ \begin{pmatrix} f_i^{**}(x_i) \\ A_i x_i \end{pmatrix} \mid x_i \in \operatorname{conv} X_i \right\}, \quad i = 1, \ldots, n,$

Primal Solutions

Primalization. We seek an explicit convex representation of z^*.

- Remember f_i and f_i^{**} match at extreme points of $\mathcal{C}_i$.
- Approximating z^* using extreme points of $\mathcal{C}_i$ thus reduces the duality gap.

Use **Frank-Wolfe** to solve a bounded equivalent of

$$\min_z \ \|z - z^*\|^2 \quad \text{subject to } z \in \mathcal{C} + \mathbb{R}_+^{m+1}. \tag{2}$$

written

$$\min_z \ \|z - z^*\|_+^2 \quad \text{subject to } z \in \mathcal{C}. \tag{3}$$

Any solution of problem (3) has optimal value zero, solves (CoP) and, **thanks to FW, writes the solution as a convex combination of extreme points of $\mathcal{C}_i$.**

Primal Solutions

LMO. Write $(\alpha_k, g^k) \in \mathbb{R} \times \mathbb{R}^m$ the gradient of the objective function.

- The linear minimization step then reads

$$s^k \in \underset{z \in \sum_{i=1}^n \mathcal{C}_i}{\operatorname{argmin}} \; z^\top (\alpha_k, g^k). \tag{4}$$

  This problem is separable

$$y_i^k \in \underset{x_i \in \operatorname{conv} X_i}{\operatorname{argmin}} \; \alpha_k f_i^{**}(x_i) + (g^k)^\top A_i x_i, \tag{5}$$

  and amounts to solving, for each $i \in \{1, \ldots, n\}$,

$$y_i^k \in \underset{x_i \in \operatorname{conv} X_i}{\operatorname{argmax}} \left(-\frac{A_i^\top g^k}{\alpha_k} \right)^\top x_i - f_i^{**}(x_i) = \partial f_i^* \left(-\frac{A_i^\top g^k}{\alpha_k} \right).$$

- The key subproblem in conditional gradient methods, namely the linear minimization oracle (4), is then **as tractable as computing the conjugates of the functions f_i.**

Primal Solutions

Trimming. Once we get a convex representation with FW, a constructive version of Carathéodory gets a representation with at most $m+1$ nontrivial convex representations, hence at most $(m+1)$ terms where f_i and f_i^{**} do not match.

Proposition [Dubois-Taine and A., 2024]

Primalization. *Suppose that X_i is convex for all $i = 1, \ldots, n$, and let v^* be the optimal value of* (CoP). *Assume that we run Frank-Wolfe for K iterations, followed by Carathéodory to trim the number of elements and set $\bar{x} \in \mathbb{R}^d$ be the final point obtained. Then $\bar{x}$ satisfies*

$$\sum_{i=1}^{n} f_i(\bar{x}_i) \leq v^* + \frac{2D_{\mathcal{C}}}{\sqrt{K+1}} + (m+1)\max_i \rho(f_i),$$

$$\left\| \sum_{i=1}^{n} A_i \bar{x}_i - b \right\|_+ \leq \frac{2D_{\mathcal{C}}}{\sqrt{K+1}}.$$

Outline

- The Shapley-Folkman Theorem
- Duality Gap Bounds
- Primalization
- **Applications**

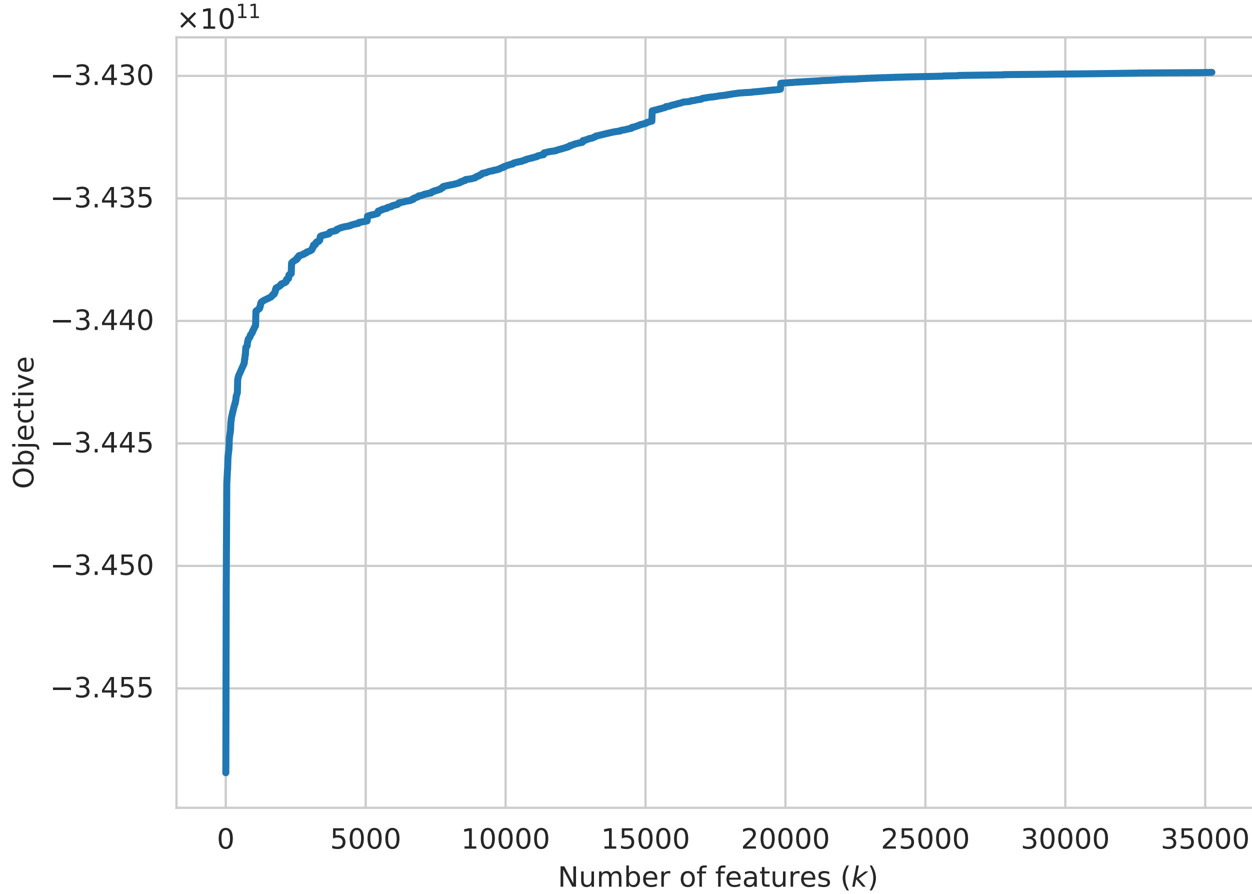
Feature Selection

- Reduce number of variables while preserving classification performance.
- Often improves test performance, especially when samples are scarce.
- Helps interpretation.

Classical examples: LASSO, ℓ_1-logistic regression, RFE-SVM, . . .

Introduction: feature selection

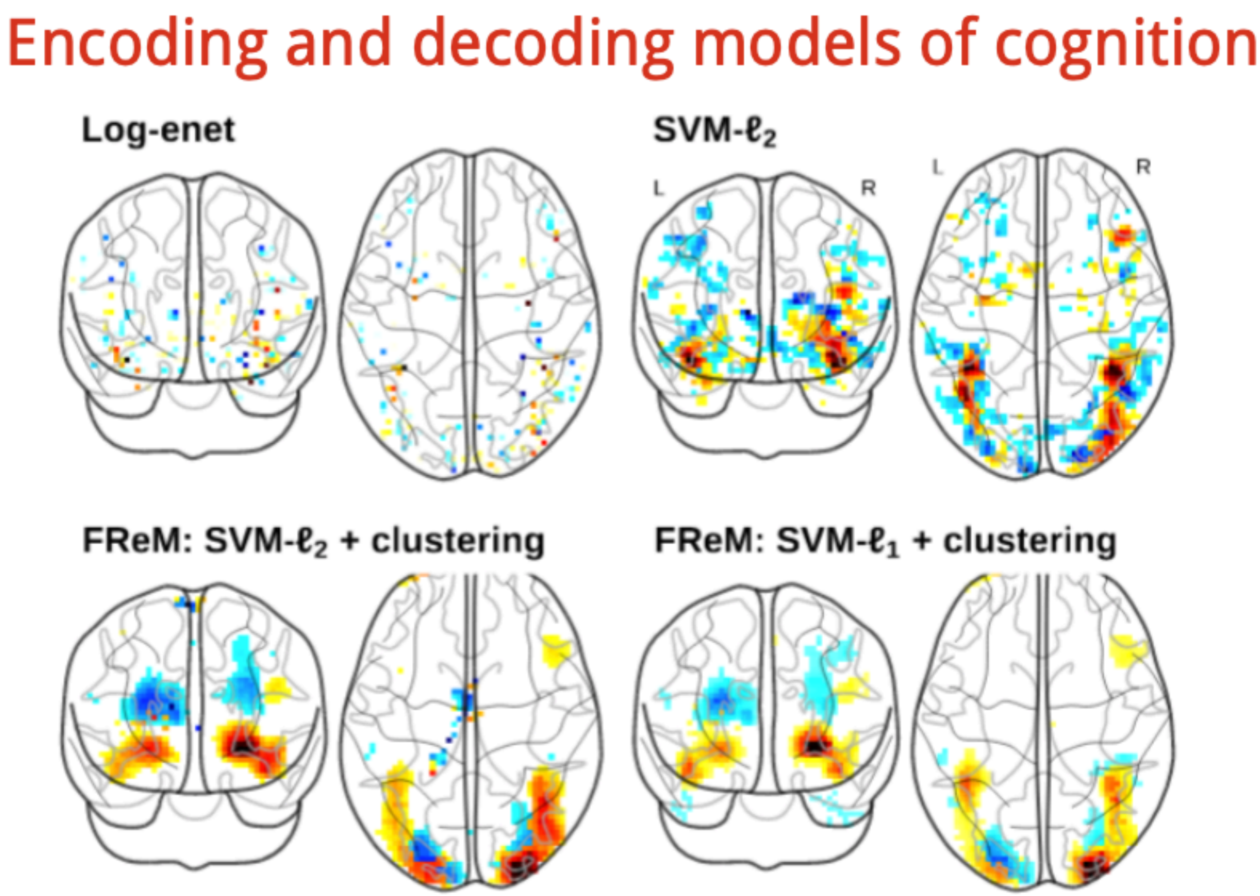
RNA classification. Find genes which best discriminate cell type (lung cancer vs control). 35238 genes, 2695 examples. [Lachmann et al., 2018]



Best ten genes: MT-CO3, MT-ND4, MT-CYB, RP11-217O12.1, LYZ, EEF1A1, MT-CO1, HBA2, HBB, HBA1.

Introduction: feature selection

Applications. Mapping brain activity by **fMRI.**



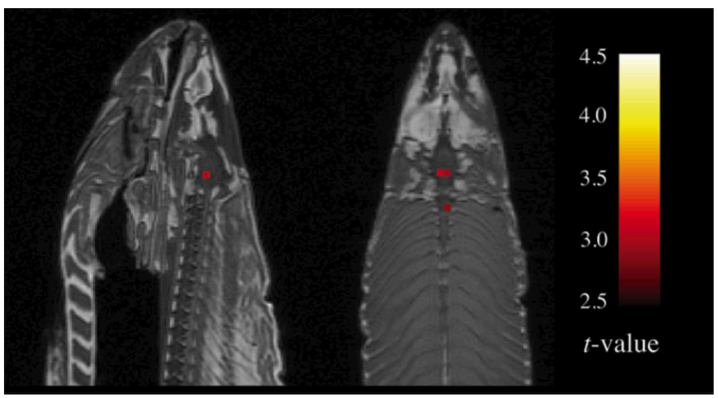
From PARIETAL team at INRIA.

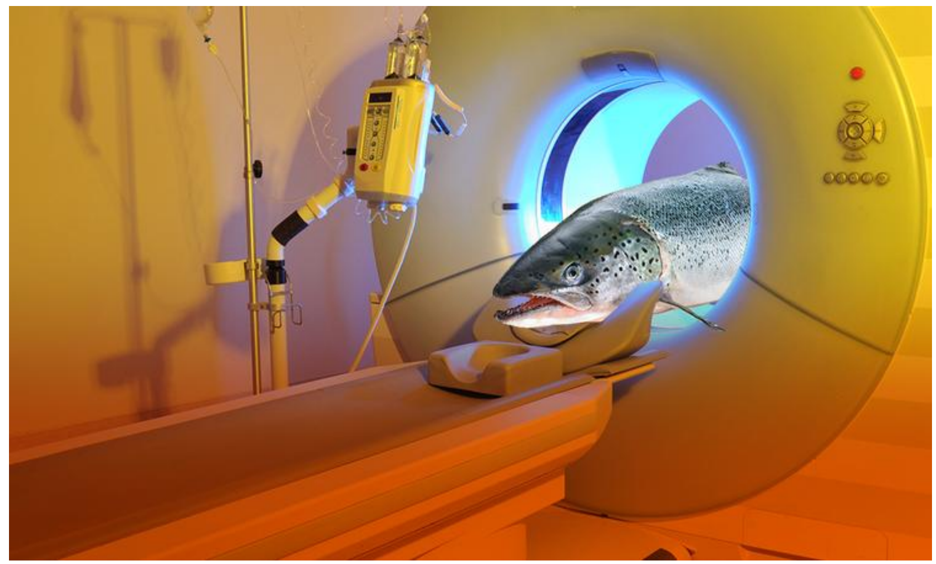
Introduction: feature selection

fMRI. Many voxels, very few samples leads to **false discoveries.**

ALEXIS MADRIGAL SCIENCE 09.18.09 05:37 PM

Scanning Dead Salmon in fMRI Machine Highlights Risk of Red Herrings





Wired article on Bennett et al. "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction" Journal of Serendipitous and Unexpected Results, 2010.

Multinomial Naive Bayse

Multinomial Naive Bayse. In the multinomial model

$$\log \mathbf{Prob}(x \mid C_\pm) = x^\top \log \theta^\pm + \log \left(\frac{(\sum_{j=1}^m x_j)!}{\prod_{j=1}^m x_j!} \right).$$

Training by maximum likelihood

$$(\theta_*^+, \theta_*^-) = \underset{\substack{\mathbf{1}^\top \theta^+ = \mathbf{1}^\top \theta^- = 1 \\ \theta^+, \theta^- \in [0,1]^m}}{\mathrm{argmax}} \; f^{+\top} \log \theta^+ + f^{-\top} \log \theta^-$$

where $f^\pm$ are sum of positive (resp. negative) feature vectors. Linear classification rule: for a given test point $x \in \mathbb{R}^m$, set

$$\hat{y}(x) = \textbf{sign}(v + w^\top x),$$

where

$$w \triangleq \log \theta_*^+ - \log \theta_*^- \quad \text{and} \quad v \triangleq \log \mathbf{Prob}(C_+) - \log \mathbf{Prob}(C_-),$$

Sparse Naive Bayse

Naive Feature Selection. Make $w \triangleq \log \theta_*^+ - \log \theta_*^-$ sparse.

Solve

$$
(\theta_*^+, \theta_*^-) = \begin{array}{ll} \text{argmax} & f^{+\top} \log \theta^+ + f^{-\top} \log \theta^- \\ \text{subject to} & \|\theta^+ - \theta^-\|_0 \leq k \\ & \mathbf{1}^\top \theta^+ = \mathbf{1}^\top \theta^- = 1 \\ & \theta^+, \theta^+ \geq 0 \end{array} \qquad \text{(SMNB)}
$$

where $k \geq 0$ is a target number of features. Features for which $\theta_i^+ = \theta_i^-$ can be discarded.

Nonconvex problem.

- Convex relaxation?
- Approximation bounds?

Sparse Naive Bayse

Convex Relaxation. The **dual is very simple.**

Sparse Multinomial Naive Bayes [Askari, A., El Ghaoui, 2019]

Let $\phi(k)$ be the optimal value of (SMNB). Then $\phi(k) \leq \psi(k)$, where $\psi(k)$ is the optimal value of the following one-dimensional convex optimization problem

$$\psi(k) := C + \min_{\alpha \in [0,1]} s_k(h(\alpha)), \qquad \text{(USMNB)}$$

where C is a constant, $s_k(\cdot)$ is the sum of the top k entries of its vector argument, and for $\alpha \in (0,1)$,

$$h(\alpha) := f_+ \circ \log f_+ + f_- \circ \log f_- - (f_+ + f_-) \circ \log(f_+ + f_-) - f_+ \log \alpha - f_- \log(1-\alpha).$$

Solved by bisection, linear complexity $O(n + k \log k)$.

Naive Feature Selection

Duality gap bound. Sparse naive Bayes reads

$$\begin{array}{lll} \mathrm{h}_P(u) = & \min_{q,r} & -f^{+\top}\log q - f^{-\top}\log r \\ & \text{subject to} & \mathbf{1}^\top q = 1 + u_1, \\ & & \mathbf{1}^\top r = 1 + u_2, \\ & & \sum_{i=1}^m \mathbf{1}_{q_i \neq r_i} \leq k + u_3 \end{array}$$

in the variables $q, r \in [0,1]^m$, where $u \in \mathbb{R}^3$. There are three constraints, two of them convex, which means $\bar{p} = (0, 0, 4)$.

Theorem [Askari, A., El Ghaoui, 2019]

NFS duality gap bounds. *Let* $\phi(k)$ *be the optimal value of* (SMNB) *and* $\psi(k)$ *that of the convex relaxation* (USMNB). *We have*

$$\psi(k-4) \leq \phi(k) \leq \psi(k),$$

for $k \geq 4$.

Sparse Programs

Sparse Programs. Low rank data and sparsity constraints

$$p_{\rm con}(k) \triangleq \min_{\|w\|_0 \leq k} \; f(Xw) + \frac{\gamma}{2}\|w\|_2^2, \tag{P-CON}$$

in the variable $w \in \mathbb{R}^m$, where $X \in \mathbb{R}^{n\times m}$ is **low rank**, $y \in \mathbb{R}^n, \gamma > 0$ and $k \geq 0$.

Penalized formulation

$$p_{\rm pen}(\lambda) \triangleq \min_{w} \; f(Xw) + \frac{\gamma}{2}\|w\|_2^2 + \lambda\|w\|_0 \tag{P-PEN}$$

in the variable $w \in \mathbb{R}^m$, where $\lambda > 0$.

Key examples: LASSO, ℓ_0-constrained logistic regression.

Naive Feature Selection

Data.

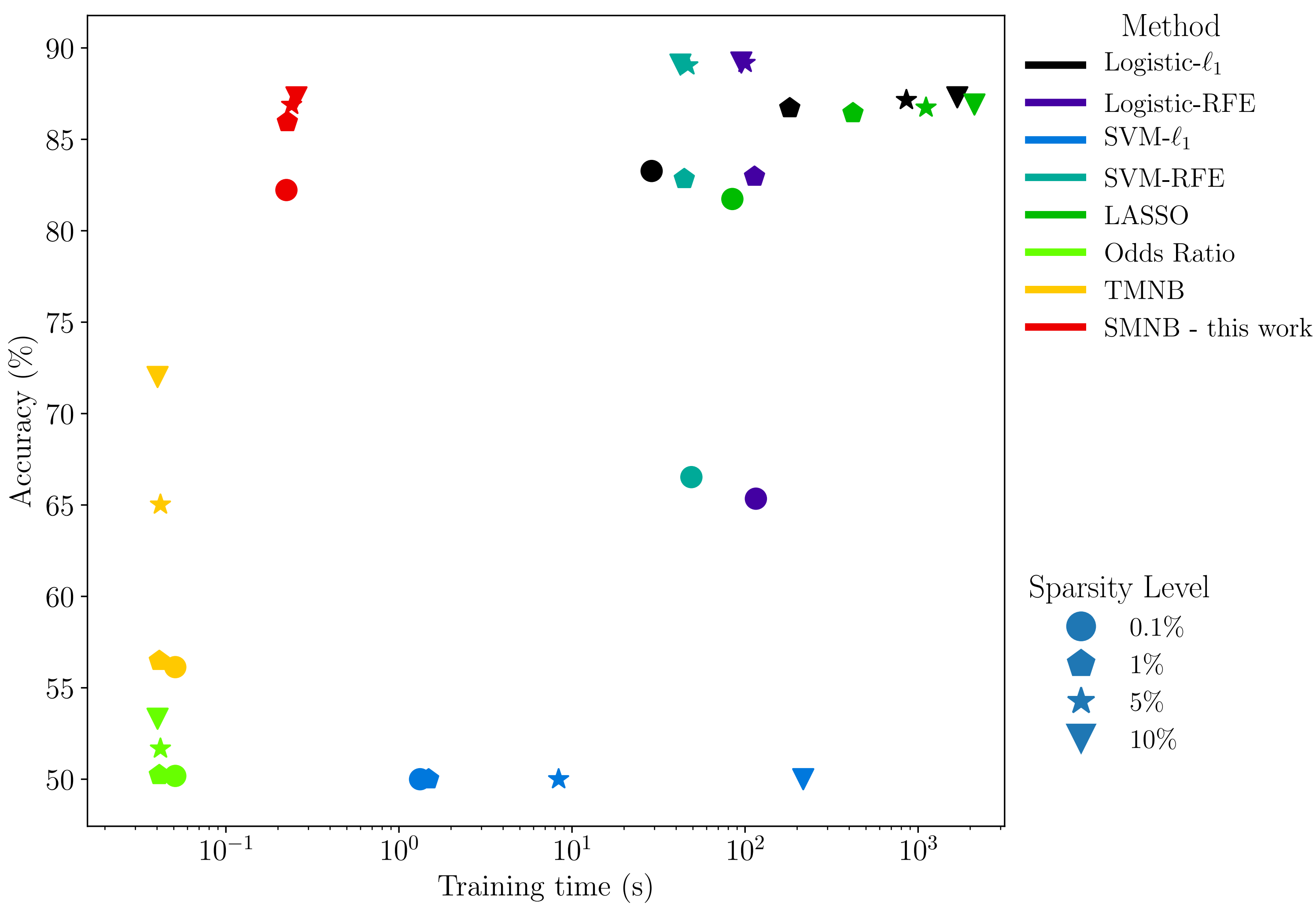
Feature Vectors	Amazon	IMDB	Twitter	MPQA	SST2
Count Vector	31,666	103,124	273,779	6,208	16,599
tf-idf	31,666	103,124	273,779	6,208	16,599
tf-idf wrd bigram	870,536	8,950,169	12,082,555	27,603	227,012
tf-idf char bigram	25,019	48,420	17,812	4838	7762

Number of features in text data sets used below.

	Amazon	IMDB	Twitter	MPQA	SST2
Count Vector	0.043	0.22	1.15	0.0082	0.037
tf-idf	0.033	0.16	0.89	0.0080	0.027
tf-idf wrd bigram	0.68	9.38	13.25	0.024	0.21
tf-idf char bigram	0.076	0.47	4.07	0.0084	0.082

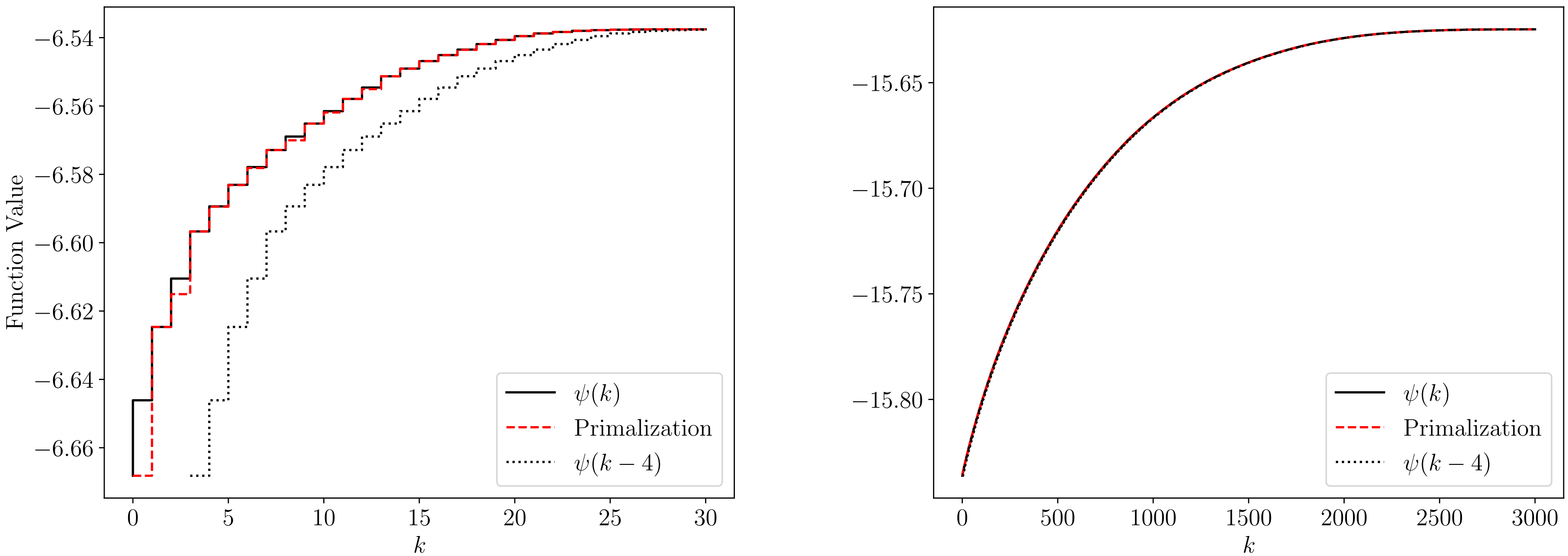
Average run time (seconds, plain Python on CPU).

Naive Feature Selection.



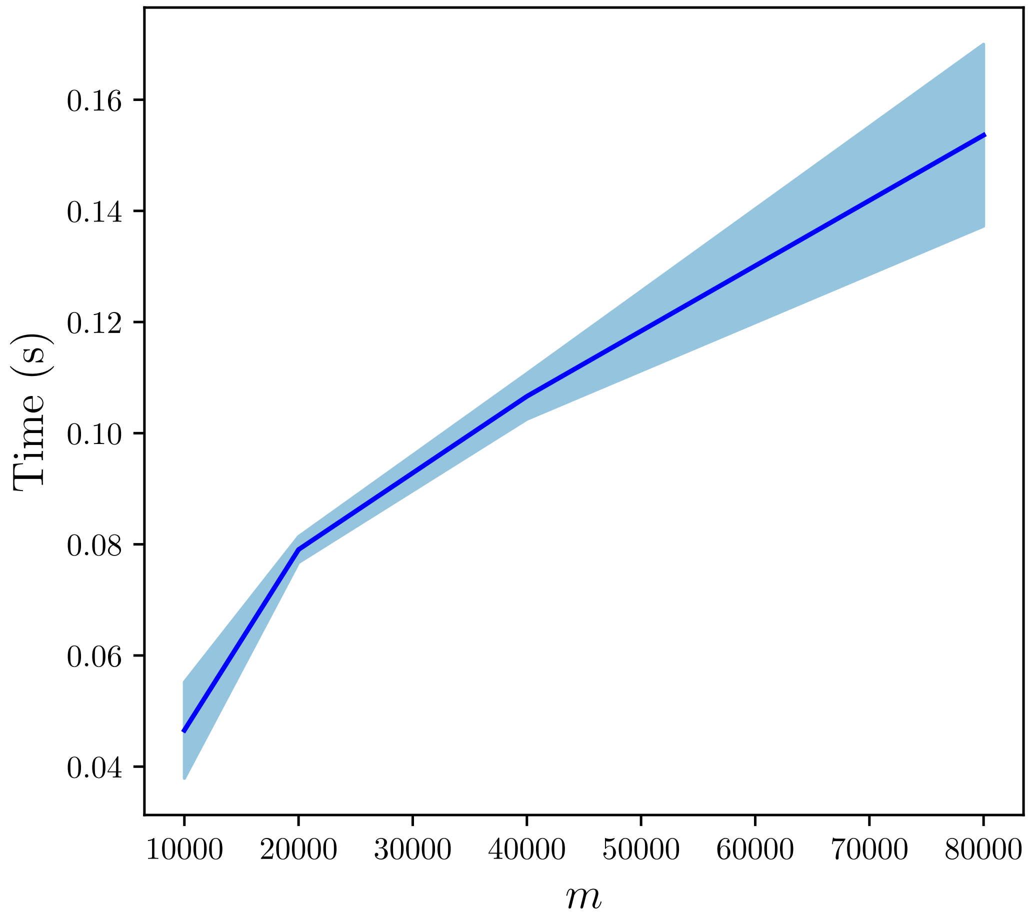
Accuracy versus run time on IMDB/Count Vector, MNB in stage two.

Naive Feature Selection.



Duality gap bound versus sparsity level for $m = 30$ (left panel) and $m = 3000$ (right panel), showing that the duality gap quickly closes as m or k increase.

Naive Feature Selection.

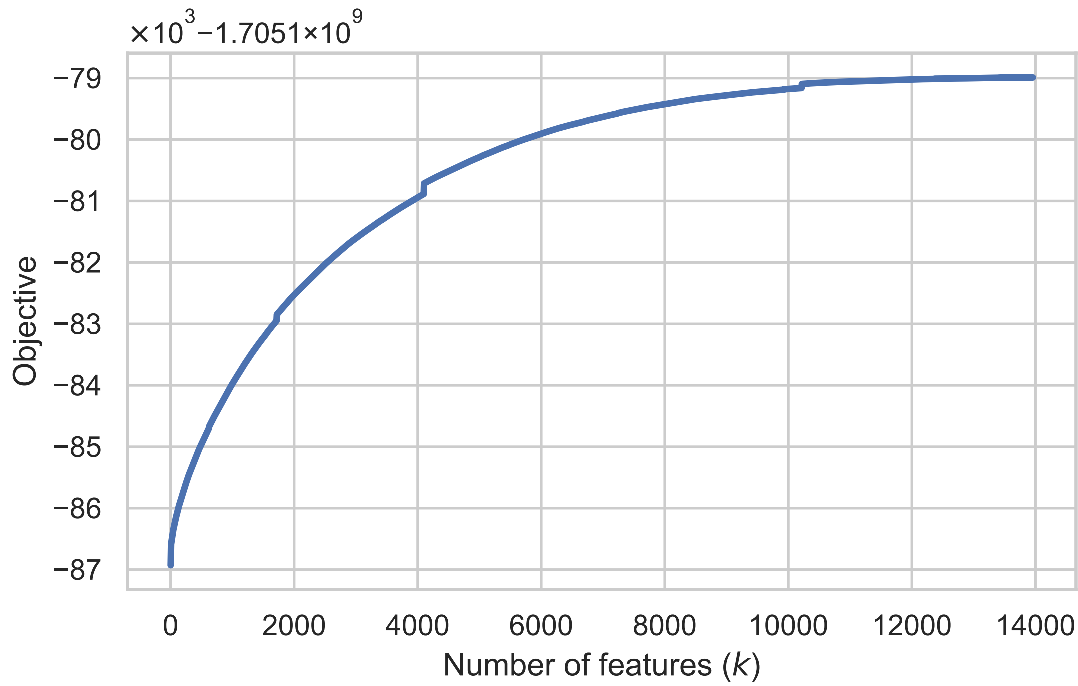


Run time with IMDB dataset/tf-idf vector data set, with increasing m, k with fixed ratio k/m, empirically showing (sub-) linear complexity.

Naive Feature Selection.

Criteo data set. Conversion logs. 45 GB, 45 million rows, 15000 columns.

- Preprocessing (NaN, encoding categorical features) takes 50 minutes.
- Computing f^+ and f^- takes 20 minutes.
- Computing the full curve below (i.e. solving 15000 problems) takes **2 minutes.**



Standard workstation, plain Python on CPU.

Conclusion

Shapley Folkman.

- Duality gap bounds for separable problems with linear constraints.
- Systematic primalization.
- Applications to Sparse Naive Bayes, LASSO, ℓ_0-logistic regression. . .

For naive Bayes, we get sparsity almost for free.

Papers: ArXiv:1905.09884 at AISTATS 2020, ArXiv:2102.06742 in *SIAM Journal on the Mathematics of Data Science*, 4 (2), pp. 514-530, 2022, and ArXiv:2406.18282, to appear in *Mathematical Programming*.

Python code: `https://github.com/aspremon/NaiveFeatureSelection`

Stable bounds on duality gap.

Active constraints. [Udell and Boyd, 2016] show that we can replace the number of contraints m by the **number of active contraints $\tilde{m}$.**

- Write the optimal set

$$X^\star = \{M_1 \times \ldots \times M_n\} \cap \{Ax \leq b\}, \quad \text{where } M_i = \mathop{\mathrm{argmin}}_{x_i \in Y_i} f_i^{**}(x_i) + \lambda^{\star T} A x_i$$

- x is an extreme point of $X^\star$ if and only if x is the only point at intersection of minimal faces F_1, F_2 of resp. $\{M_1 \times \ldots \times M_n\}$ and $\{Ax \leq b\}$ containing x [Dubins, 1962, Th. 5.1], [Udell and Boyd, 2016, Lem. 3].

- This means that $\dim F_1 + \dim F_2 \leq d$ with $d - \tilde{m} \leq \dim F_2$, so $\dim F_1 \leq \tilde{m}$.

- As faces of Cartesian products are Cartesian products of faces, the sum of dimensions of the faces of M_i containing x_i is smaller than $\tilde{m}$, hence at least $n - \tilde{m}$ points x_i of these faces are extreme points where $f_i^{**}(x_i) = f_i(x_i)$.

*

References

Jean-Pierre Aubin and Ivar Ekeland. Estimates of the duality gap in nonconvex optimization. *Mathematics of Operations Research*, 1(3): 225–245, 1976.

Lester E Dubins. On extreme points of convex sets. *Journal of Mathematical Analysis and Applications*, 5(2):237–244, 1962.

Benjamin Dubois-Taine and Alexandre d'Aspremont. Frank-wolfe meets shapley-folkman: a systematic approach for solving nonconvex separable problems with linear constraints. *arXiv preprint arXiv:2406.18282*, 2024.

Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.

William R Emerson and Frederick P Greenleaf. Asymptotic behavior of products cp= c+...+c in locally compact abelian groups. *Transactions of the American Mathematical Society*, 145:171–204, 1969.

Matthieu Fradelizi, Mokshay Madiman, Arnaud Marsiglietti, and Artem Zvavitch. The convexification effect of minkowski summation. *Preprint*, 2017.

Alexander Lachmann, Denis Torre, Alexandra B Keenan, Kathleen M Jagodnik, Hoyjin J Lee, Lily Wang, Moshe C Silverstein, and Avi Ma'ayan. Massive mining of publicly available rna-seq data from human and mouse. *Nature communications*, 9(1):1366, 2018.

Claude Lemaréchal and Arnaud Renaud. A geometric study of duality gaps, with applications. *Mathematical Programming*, 90(3):399–427, 2001.

Ross M Starr. Quasi-equilibria in markets with non-convex preferences. *Econometrica: journal of the Econometric Society*, pages 25–38, 1969.

Madeleine Udell and Stephen Boyd. Bounding duality gap for separable problems with linear constraints. *Computational Optimization and Applications*, 64(2):355–378, 2016.