

# Feature Selection & the Shapley-Folkman Theorem.

**Alexandre d'Aspremont,**

*CNRS & D.I., École Normale Supérieure.*

With Armin Askari, Laurent El Ghaoui (UC Berkeley)  
and Quentin Rebjock (EPFL)

# Introduction

---

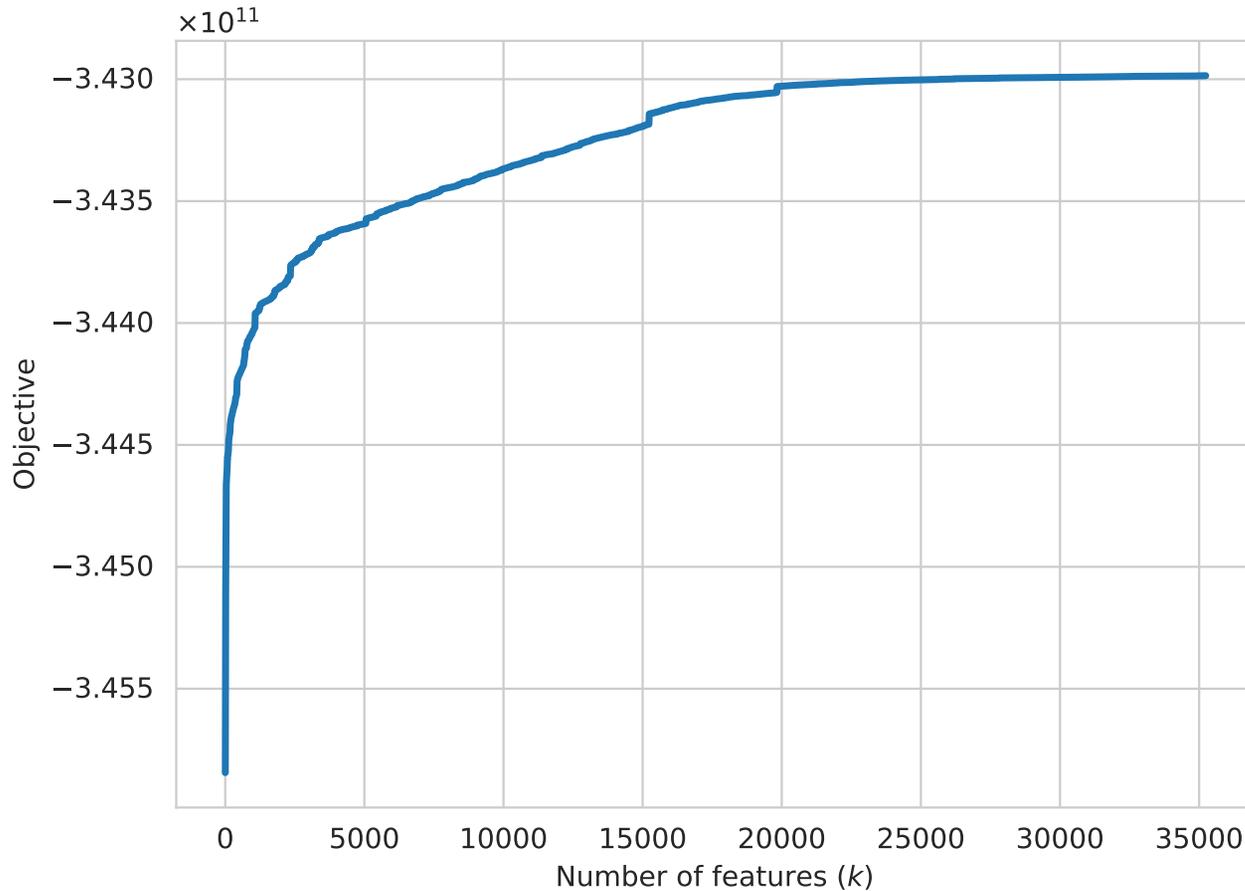
## Feature Selection.

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

**Classical examples:** LASSO,  $\ell_1$ -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]



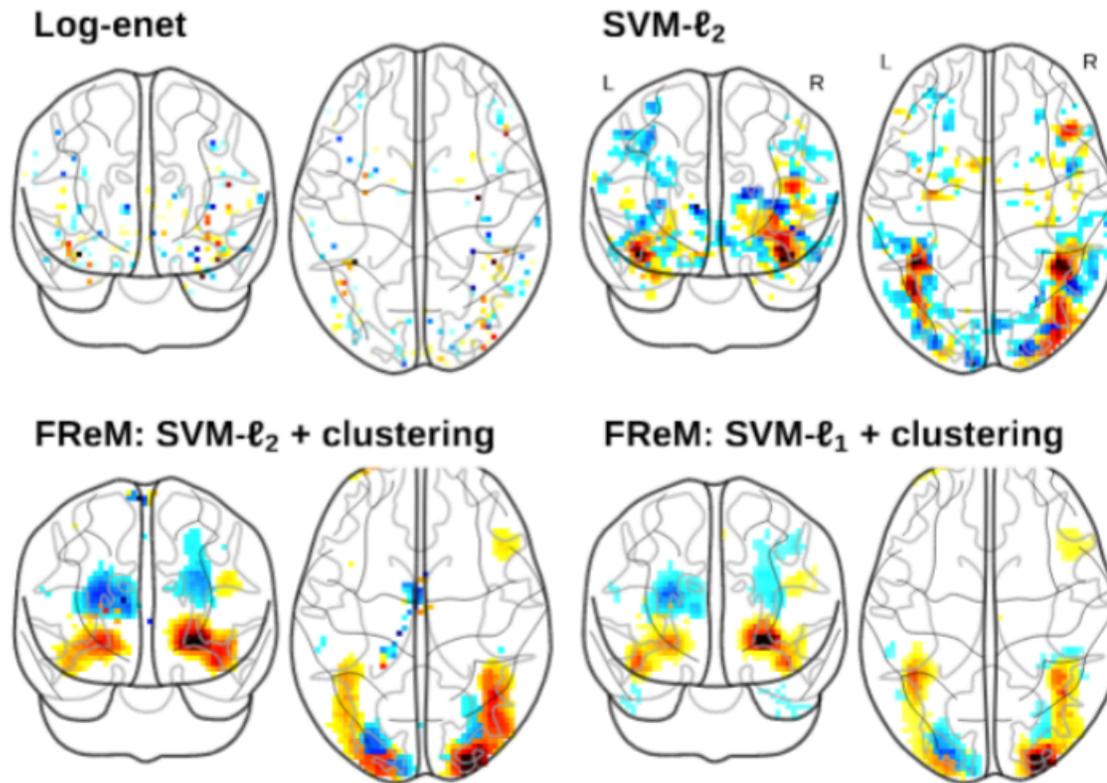
**Best ten genes:** MT-CO3, MT-ND4, MT-CYB, RP11-217012.1, LYZ, EEF1A1, MT-CO1, HBA2, HBB, HBA1.

# Introduction: feature selection

---

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

## Encoding and decoding models of cognition



From PARIETAL team at INRIA.

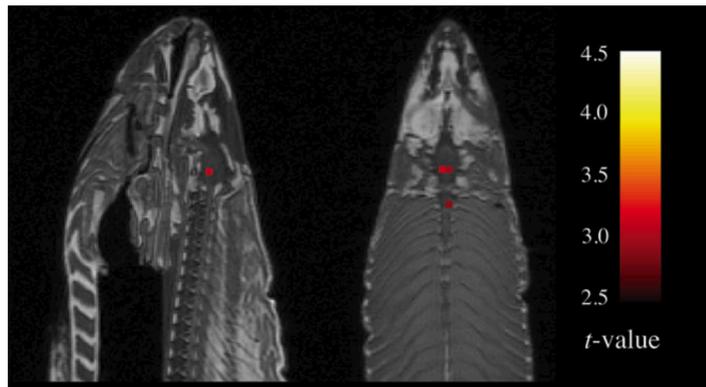
# Introduction: feature selection

---

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.

# Introduction: linear models

---

**Linear models.** Select features from large weights  $w$ .

- LASSO solves  $\min_w \|Xw - y\|_2^2 + \lambda\|w\|_1$  with linear prediction given by  $w^T x$ .
- Linear SVM, solves  $\min_w \sum_i \max\{0, 1 - y_i w^T x_i\} + \lambda\|w\|_2^2$  with linear classification rule  $\text{sign}(w^T x)$ .
- $\ell_1$ -logistic regression, etc.

**In practice.**

- Relatively **high complexity** on very large-scale data sets.
- Recovery results require **uncorrelated features** (incoherence, RIP, etc.).
- Cheaper featurewise methods (ANOVA, TF-IDF, etc.) have relatively poor performance.

# Outline

---

- **Sparse Naive Bayes**
- The Shapley-Folkman theorem
- Duality gap bounds
- Numerical performance

# Multinomial Naive Bayes

---

**Multinomial Naive Bayes.** 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

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

Linear classification rule: for a given test point  $x \in \mathbb{R}^m$ , set

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

where

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

# Sparse Naive Bayes

---

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

Solve

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

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 Bayes

---

**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

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

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  $O(n + k \log k)$ . **Approximation bounds?**

# Outline

---

- Sparse Naive Bayes
- **The Shapley-Folkman theorem**
- Duality gap bounds
- Numerical performance

# Shapley-Folkman Theorem

---

**Minkowski sum.** Given sets  $X, Y \subset \mathbb{R}^d$ , we have

$$X + Y = \{x + y : x \in X, y \in Y\}$$



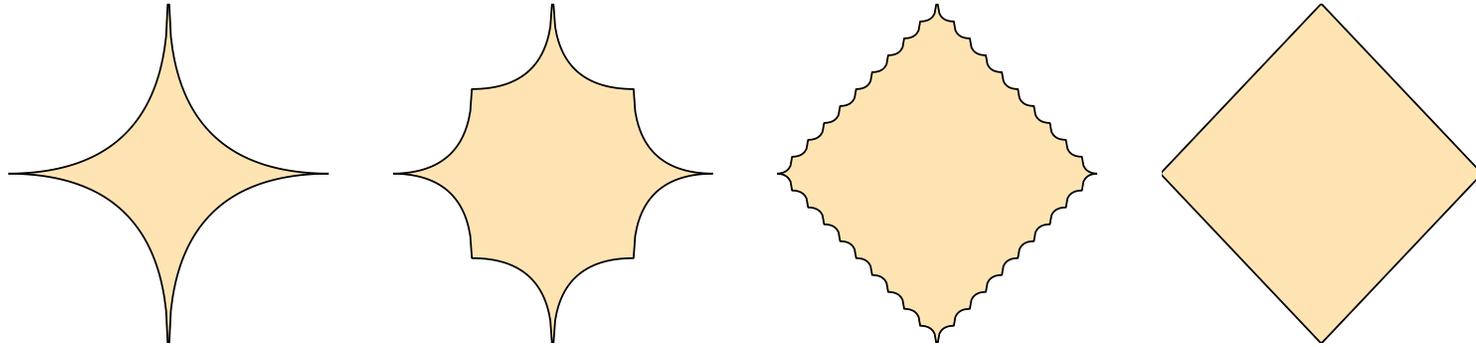
(CGAL User and Reference Manual)

**Convex hull.** Given subsets  $V_i \subset \mathbb{R}^d$ , we have

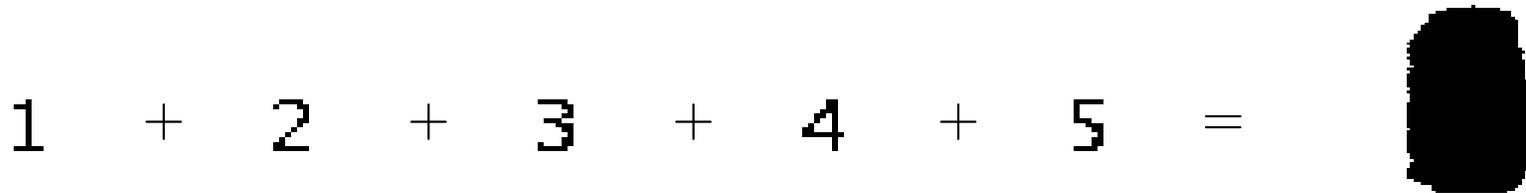
$$\text{Co} \left( \sum_i V_i \right) = \sum_i \text{Co}(V_i)$$

# Shapley-Folkman Theorem

---



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



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

# Shapley-Folkman Theorem

---

## Shapley-Folkman Theorem [Starr, 1969]

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

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

then

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

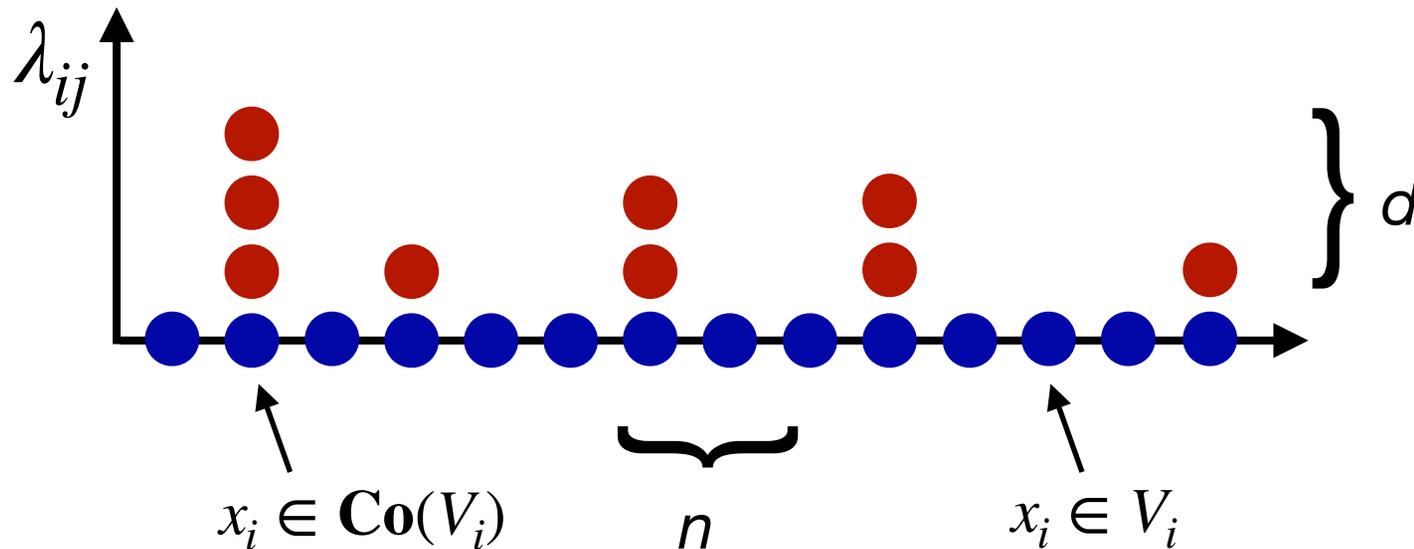
where  $|\mathcal{S}_x| \leq d$ .

# Shapley-Folkman Theorem

**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  $\lambda_{ij}$  controls gap with convex hull.**

# Shapley-Folkman: geometric consequences

---

## Consequences.

- If the sets  $V_i \subset \mathbb{R}^d$  are uniformly bounded with  $\text{rad}(V_i) \leq R$ , 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  $\text{rad}(V) = \inf_{x \in V} \sup_{y \in V} \|x - y\|$ .

- In particular, when  $d$  is fixed and  $n \rightarrow \infty$

$$\left( \frac{\sum_{i=1}^n V_i}{n} \right) \rightarrow \mathbf{Co} \left( \frac{\sum_{i=1}^n V_i}{n} \right)$$

in the Hausdorff metric with rate  $O(1/n)$ .

- Holds for many other nonconvexity measures [Fradelizi et al., 2017].

# Outline

---

- Sparse Naive Bayes
- The Shapley-Folkman theorem
- **Duality gap bounds**
- Numerical performance

# Nonconvex Optimization

---

Separable nonconvex problem. Solve

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

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**,

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

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

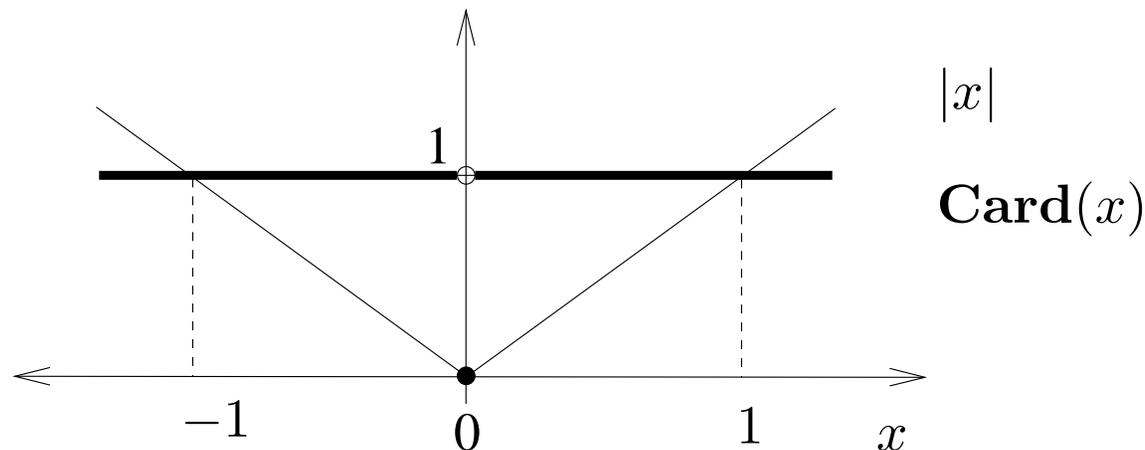
# Nonconvex Optimization

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

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

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

Example.



The  $l_1$  norm is the convex envelope of  $\text{Card}(x)$  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

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

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

Affine constraints means **(P) and (CoP) have the same dual** [Lemaréchal and Renaud, 2001, Th. 2.11], given by

$$\sup_{\lambda \geq 0} \Psi(\lambda) \tag{D}$$

in the variable  $\lambda \in \mathbb{R}^m$ . Roughly speaking, **if  $\mathcal{G}^{**} = \mathcal{G}$ , no duality gap** in (P).

# Nonconvex Optimization

---

**Epigraph & duality gap.** Define

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

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

- The epigraph  $\mathcal{G}_r^{**}$  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 at  $x \in \mathcal{G}_r^{**}$  shows  $f^{**}(x_i) = f(x_i)$  for **all but at most  $m + 1$  terms in the objective.**
- As  $n \rightarrow \infty$ , with  $m/n \rightarrow 0$ ,  $\mathcal{G}_r$  gets closer to its convex hull  $\mathcal{G}_r^{**}$  and the **duality gap becomes negligible.**

# Bound on duality gap

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  $x^* \in \mathbb{R}^d$  at which the primal optimal value of (CoP) is attained, such that

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

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

# Bound on duality gap

---

**General result.** Consider the separable nonconvex problem

$$\begin{aligned} h_P(u) := & \min. && \sum_{i=1}^n f_i(x_i) \\ & \text{s.t.} && \sum_{i=1}^n g_i(x_i) \leq b + u \end{aligned} \quad (\text{P})$$

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  $i = 1, \dots, n, j = 1, \dots, m$ . Let

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

then

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

where  $h_P(u)^{**}$  is the optimal value of the dual to (P).

# Naive Feature Selection

**Duality gap bound.** Sparse naive Bayes reads

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

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  $k \geq 4$ .

# Naive Feature Selection

---

**Primalization.** Given  $\alpha_*$  solving (USMNB), reconstruct point  $(\theta^+, \theta^-)$ .

For  $\alpha^*$  optimal for (USMNB), let  $\mathcal{I}$  be complement of the set of indices corresponding to the top  $k$  entries of  $h(\alpha_*)$ , set  $B_{\pm} := \sum_{i \notin \mathcal{I}} f_i^{\pm}$ , and

$$\theta_{*i}^+ = \theta_{*i}^- = \frac{f_i^+ + f_i^-}{\mathbf{1}^\top (f^+ + f^-)}, \quad \forall i \in \mathcal{I}, \quad \theta_{*i}^{\pm} = \frac{B_+ + B_-}{B_{\pm}} \frac{f_i^{\pm}}{\mathbf{1}^\top (f^+ + f^-)}, \quad \forall i \notin \mathcal{I}.$$

In all but pathological scenarios,  $k$  largest coefficients in (USMNB) give **support** of the solution.

# Outline

---

- Sparse Naive Bayes
- Approximation bounds & the Shapley-Folkman theorem
- **Numerical performance**

# Naive Feature Selection

---

## Data.

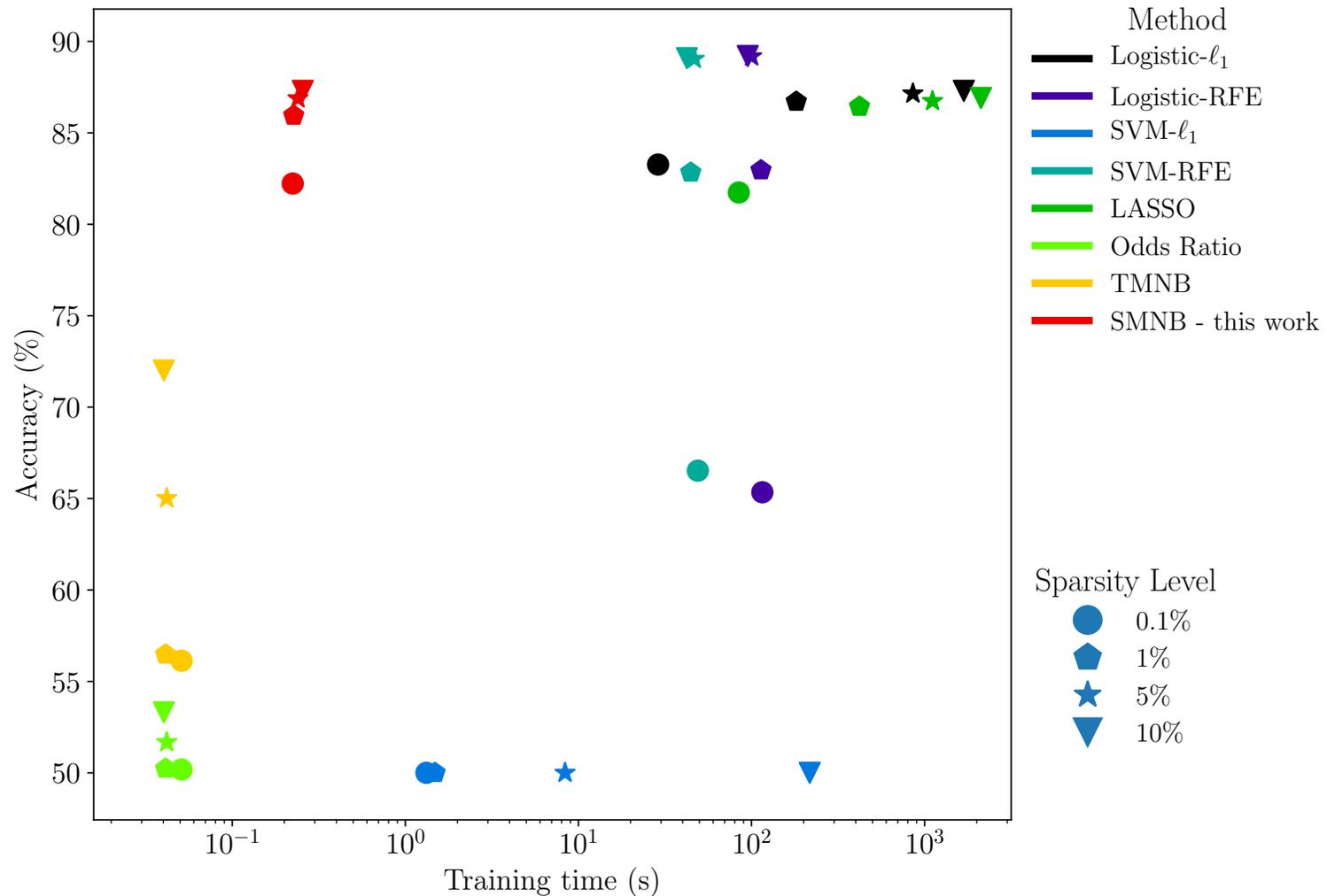
FEATURE VECTORS	AMAZON	IMDB	TWITTER	MPQA	SST2
COUNT VECTOR	31,666	103,124	273,779	6,208	16,599
TF-IDF	31,666	103,124	273,779	6,208	16,599
TF-IDF WRD BIGRAM	870,536	8,950,169	12,082,555	27,603	227,012
TF-IDF CHAR BIGRAM	25,019	48,420	17,812	4838	7762

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

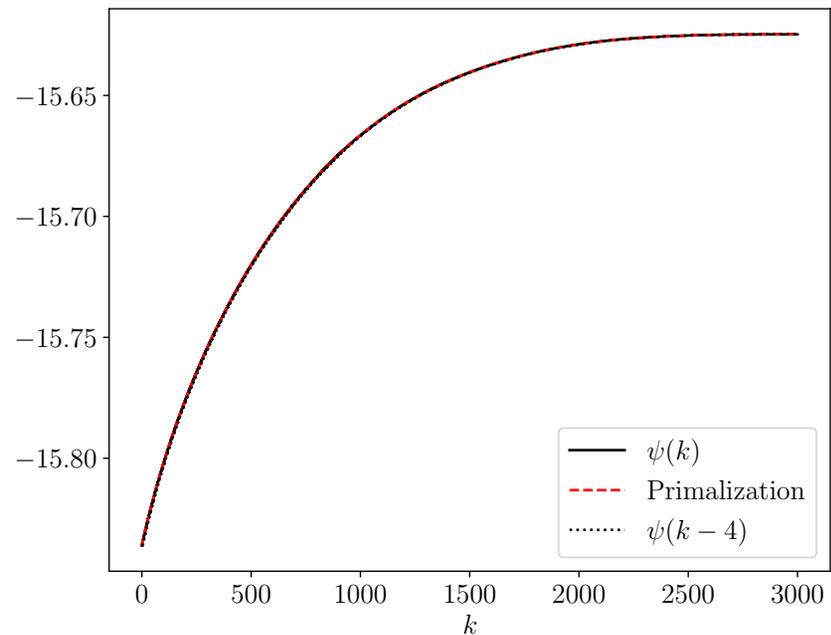
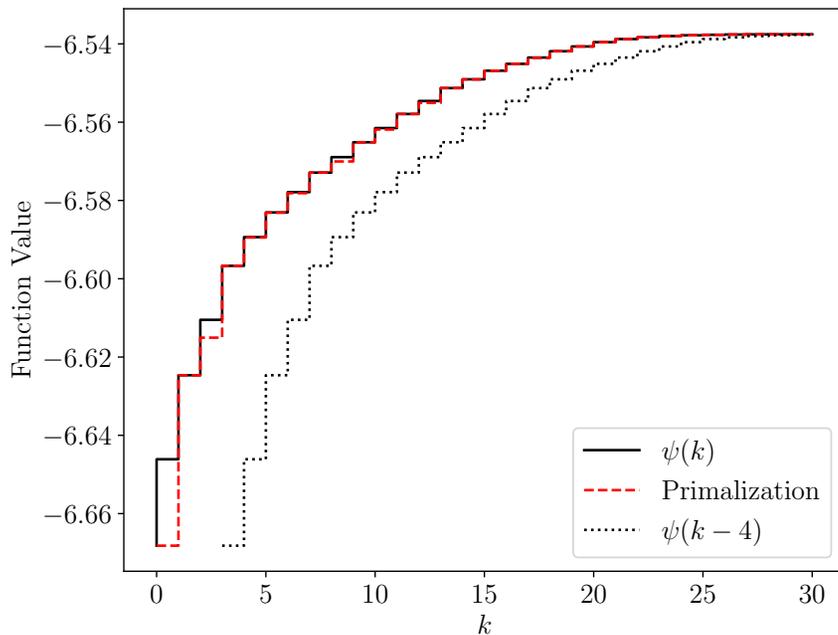
Average run time (seconds, plain Python on CPU).

# Naive Feature Selection.



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

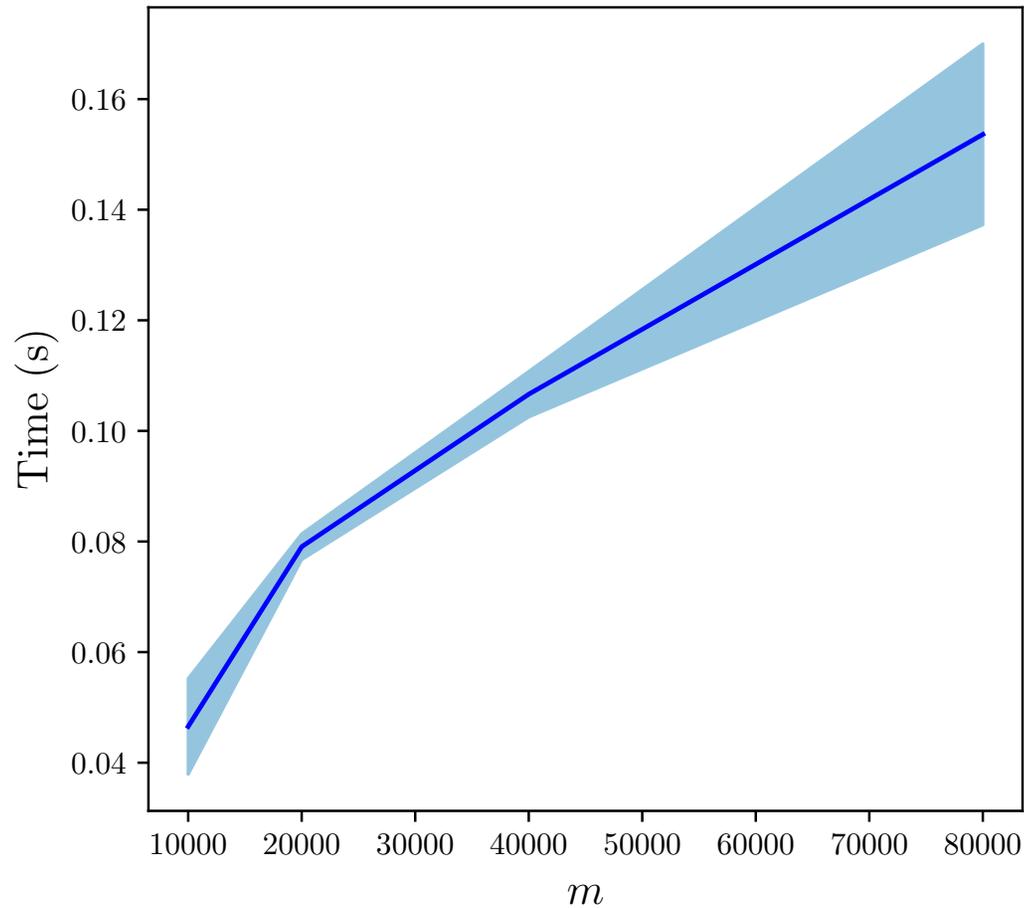
# Naive Feature Selection.



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.

---



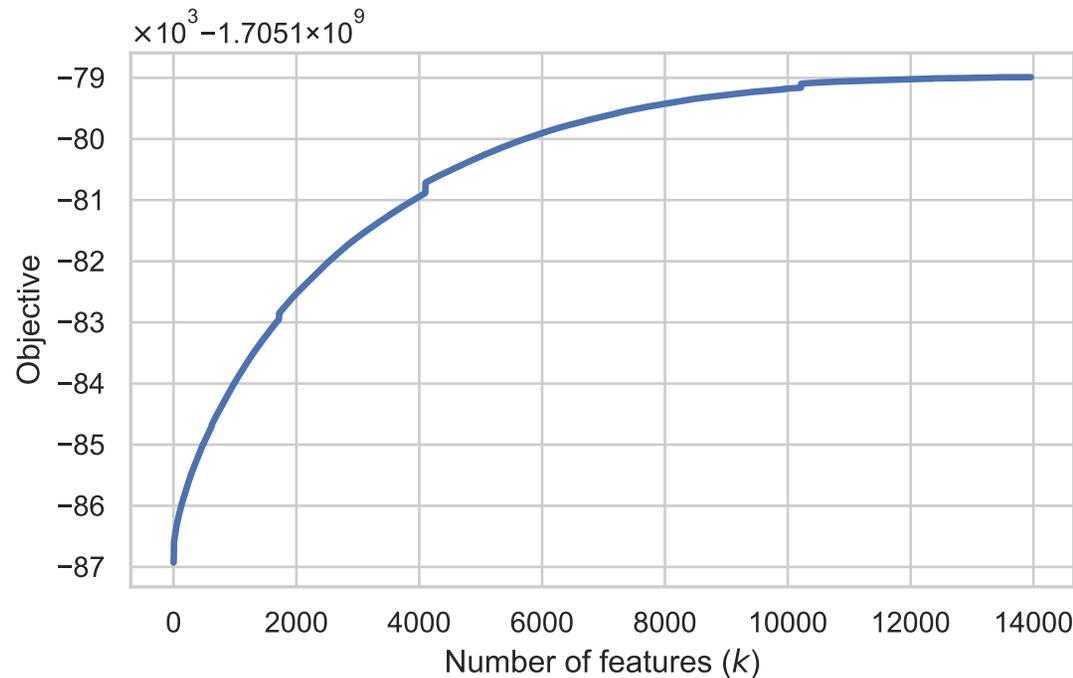
Run time with IMDB dataset/tf-idf vector data set, with increasing  $m, k$  with fixed ratio  $k/m$ , empirically showing (sub-) linear complexity.

# 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

---

## Naive Feature Selection.

**For naive Bayes, we get sparsity almost for free.**

- Linear complexity.
- Nearly tight convex relaxation.
- Feature selection performance comparable to LASSO or  $\ell_1$  logistic regression, but NFS is  $100\times$  faster. . .
- Requires no RIP assumption (only the naive one behind NB).

<https://github.com/aspremon/NaiveFeatureSelection>



---

## References

- Jean-Pierre Aubin and Ivar Ekeland. Estimates of the duality gap in nonconvex optimization. *Mathematics of Operations Research*, 1(3): 225–245, 1976.
- Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.
- 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.