Sparsity, 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

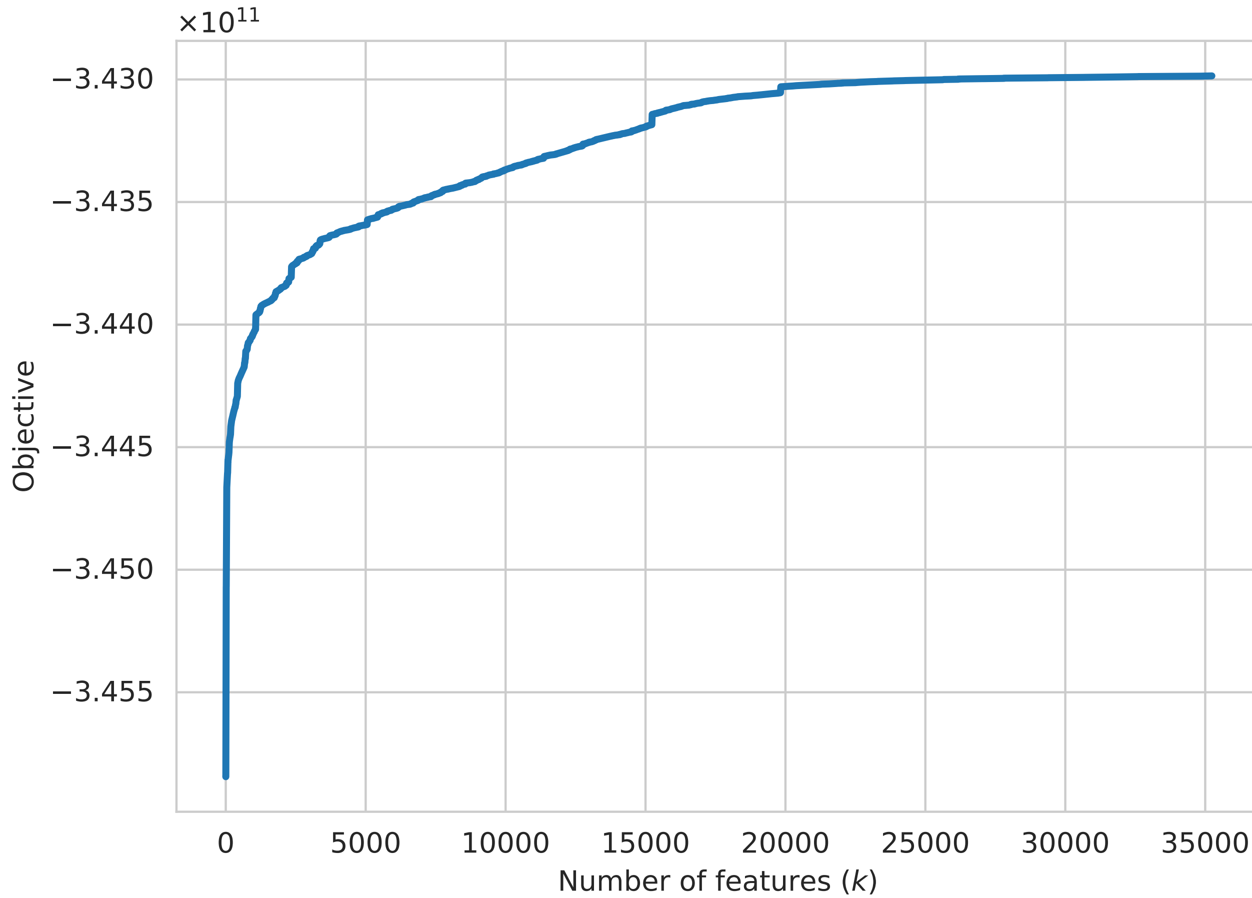
Feature Selection.

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

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

Introduction: feature selection

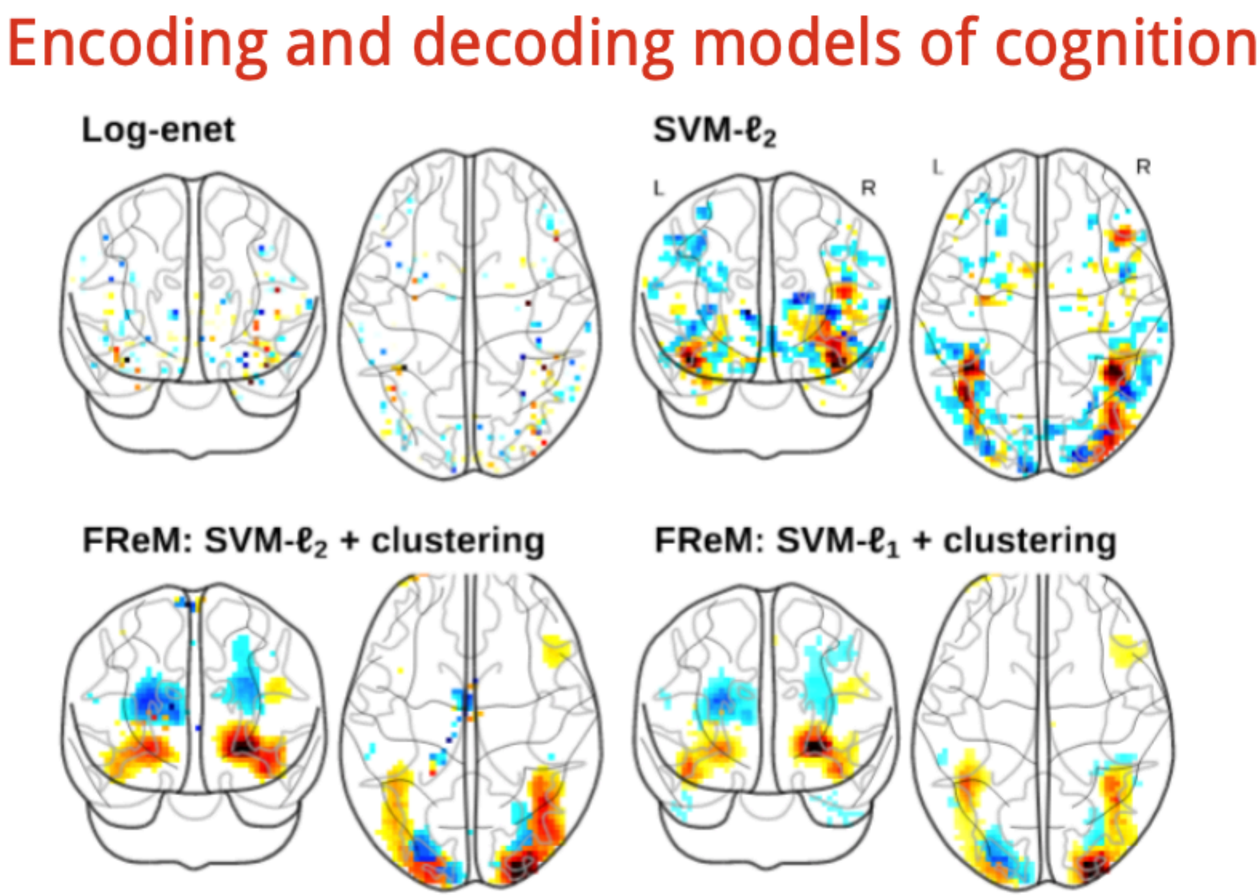
RNA classification. Find genes which best discriminate cell type (lung cancer vs control). 35238 genes, 2695 examples. [Lachmann et al., 2018]



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.**



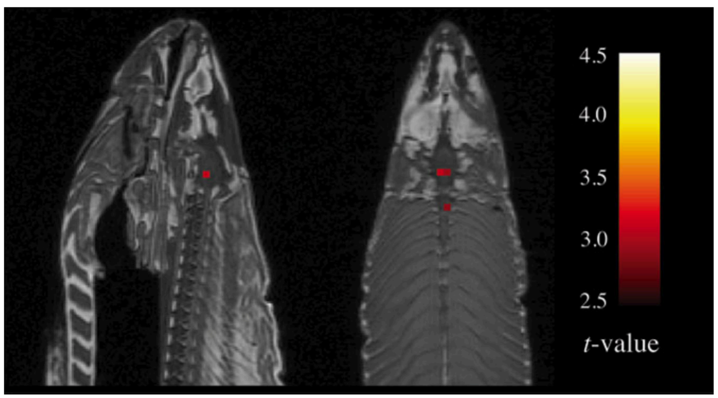
From PARIETAL team at INRIA.

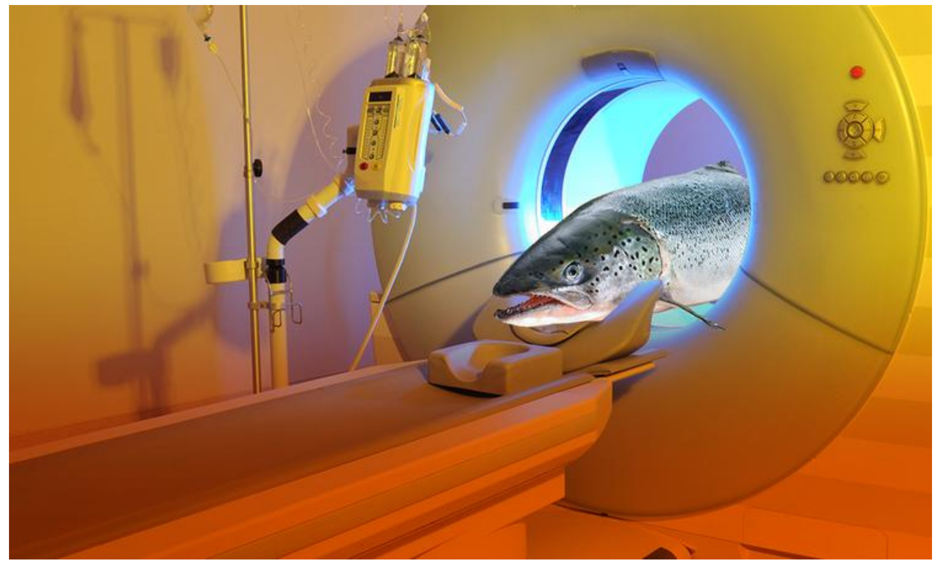
Introduction: feature selection

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.

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 $\mathrm{sign}(w^T x)$.

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
- Other Applications
- Numerical Performance

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

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

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

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

where $k \ge 0$ 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 $O(n + k \log k)$. **Approximation bounds?**

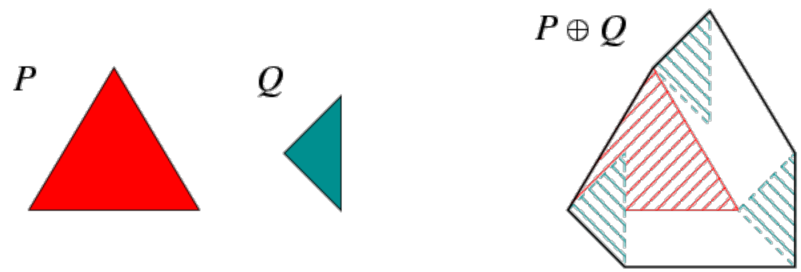
Outline

- Sparse Naive Bayes
- **The Shapley-Folkman Theorem**
- Duality Gap Bounds
- Other Applications
- Numerical Performance

Shapley-Folkman Theorem

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

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

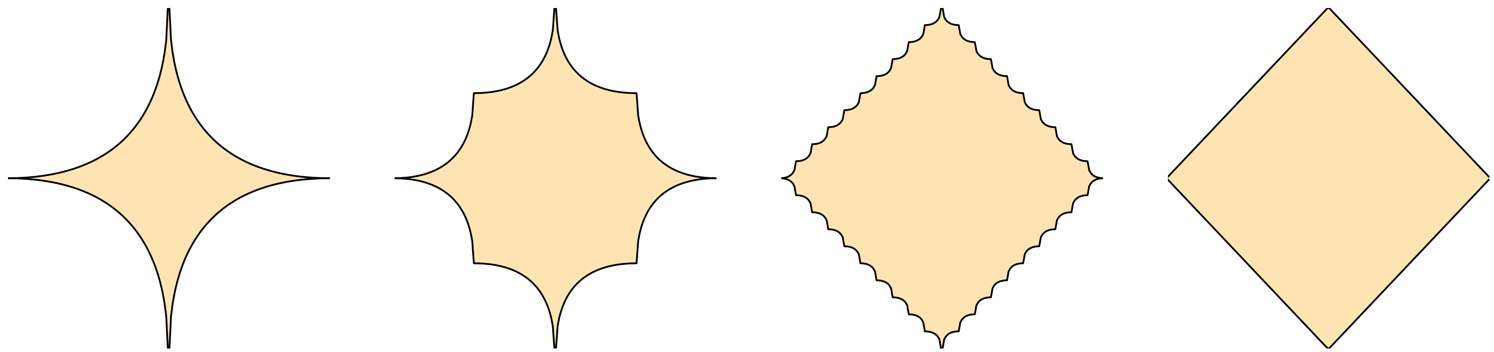


(CGAL User and Reference Manual)

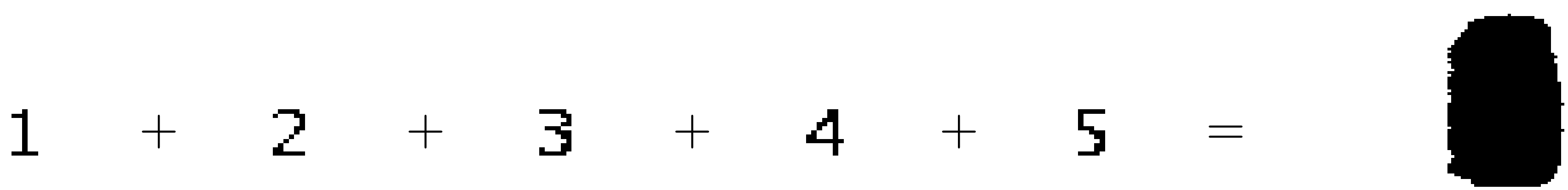
Convex hull. Given subsets $V_i \subset \mathbb{R}^d$, we have

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

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]

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

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

then

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

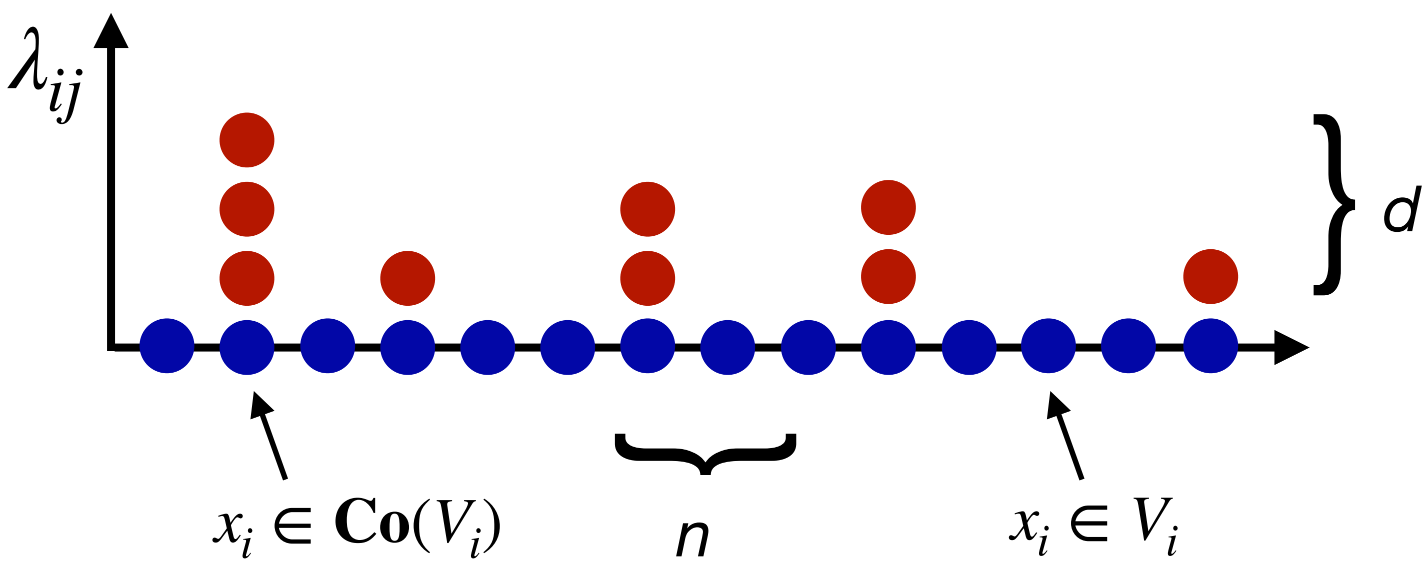
for some $|\mathcal{S}| \leq d$.

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 λ_{ij} controls gap with convex hull.

Shapley-Folkman: geometric consequences

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

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

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

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**
- Other Applications
- Numerical Performance

Nonconvex Optimization

Separable nonconvex problem. Solve

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

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

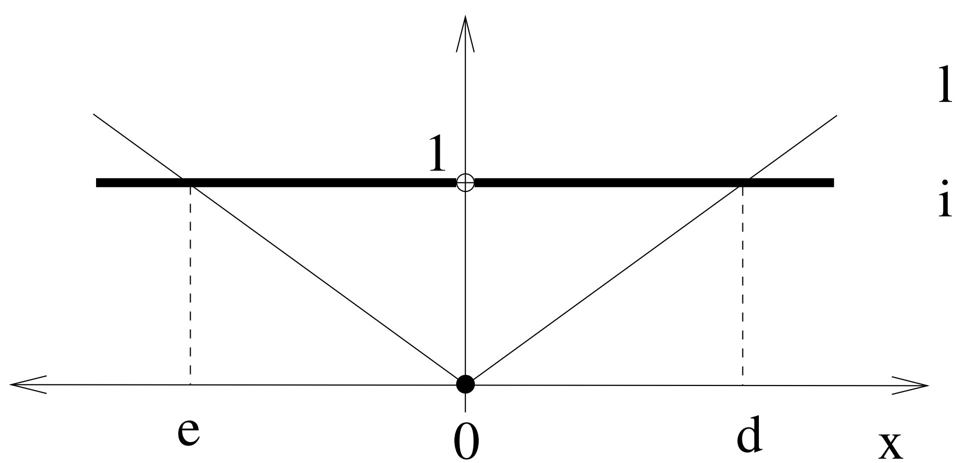
Nonconvex Optimization

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

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

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 $[-1, 1]$.

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 $\mathcal{G}_r^{**}$ 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 $n \to \infty$, with $m/n \to 0$, the **epigraph** $\mathcal{G}_r$ gets closer to $\mathcal{G}_r^{**}$, i.e. **closer to being convex**, and the **duality gap becomes negligible.**

Bound on duality gap

General result. Consider the separable nonconvex problem

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

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*

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

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)*.*

Naive Feature Selection

Duality gap bound. Sparse naive Bayes reads

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

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 $k \geq 4$.

Primalization is tricky, cf. paper. . .

Outline

- Sparse Naive Bayes
- The Shapley-Folkman Theorem
- Duality Gap Bounds
- **Other Applications**
- Numerical Performance

Sparse Programs

Problems with **low rank data and sparsity constraints**

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

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

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

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

Key examples: LASSO, ℓ_0-constrained logistic regression.

Convex Relaxation

The **bidual** of (P-CON) is written

$$p_{\text{con}}^{**}(k) = \min_{v,u \in [0,1]^m} f(XD(u)v) + \frac{\gamma}{2} v^\top D(u)v \;:\; \mathbf{1}^\top u \leq k \qquad \text{(BD-CON)}$$

Non-convex, but setting $\tilde{v} = D(u)v$ equivalent to

$$p_{\text{con}}^{**}(k) = \min_{\tilde{v},u \in [0,1]^m} f(X\tilde{v}) + \frac{\gamma}{2} \tilde{v} D(u)^\dagger \tilde{v} \;:\; \mathbf{1}^\top u \leq k \qquad (1)$$

in the variables $\tilde{v}, u \in \mathbb{R}^m$, where $\tilde{v}^\top D(u)^\dagger \tilde{v}$ is **jointly convex** in $(\tilde{v}, u)$ (second order cone constraint).

This is the **interval relaxation** of the ℓ_0 sparsity constraint.

Duality Gap Bounds

Proposition

Gap Bounds. *Suppose $X = U_r \Sigma_r V_r^\top$ is a compact, rank-r SVD decomposition of X. From a solution (v^*, u^*) of* (BD-CON) *with objective value t^*, with probability one, we can construct a point with at most $k + r + 2$ nonzero coefficients and objective value OPT satisfying*

$$p_{\text{con}}(k + r + 2) \leq OPT \leq p^{**}_{\text{con}}(k) \leq p_{\text{con}}(k) \qquad \text{(Gap-Bound)}$$

by solving a linear program written

$$\begin{array}{ll} \text{minimize} & c^\top u \\ \text{subject to} & f(U_r z^*) + \sum_{i=1}^m u_i \frac{\gamma}{2} {v_i^*}^2 = t^* \\ & \sum_{i=1}^m u_i \leq k \\ & \sum_{i=1}^m u_i \ell_i v_i^* = z^* \\ & u \in [0,1]^m \end{array} \qquad (2)$$

in the variable $u \in \mathbb{R}^m$ where $c \sim \mathcal{N}(0, I_m)$, $z^* = \Sigma_r V_r^\top D(u^*) v^*$.

Duality Gap Bounds

LASSO vs. interval.

Optimality

- Interval: only need low rank
- LASSO: need RIP, incoherence

Support Recovery

- Interval: need low rank + RIP
- LASSO: need RIP, incoherence

Both have similar computational cost.

Outline

- Sparse Naive Bayes
- The Shapley-Folkman Theorem
- Duality Gap Bounds
- Other Applications
- **Numerical Performance**

Naive Feature Selection

Data.

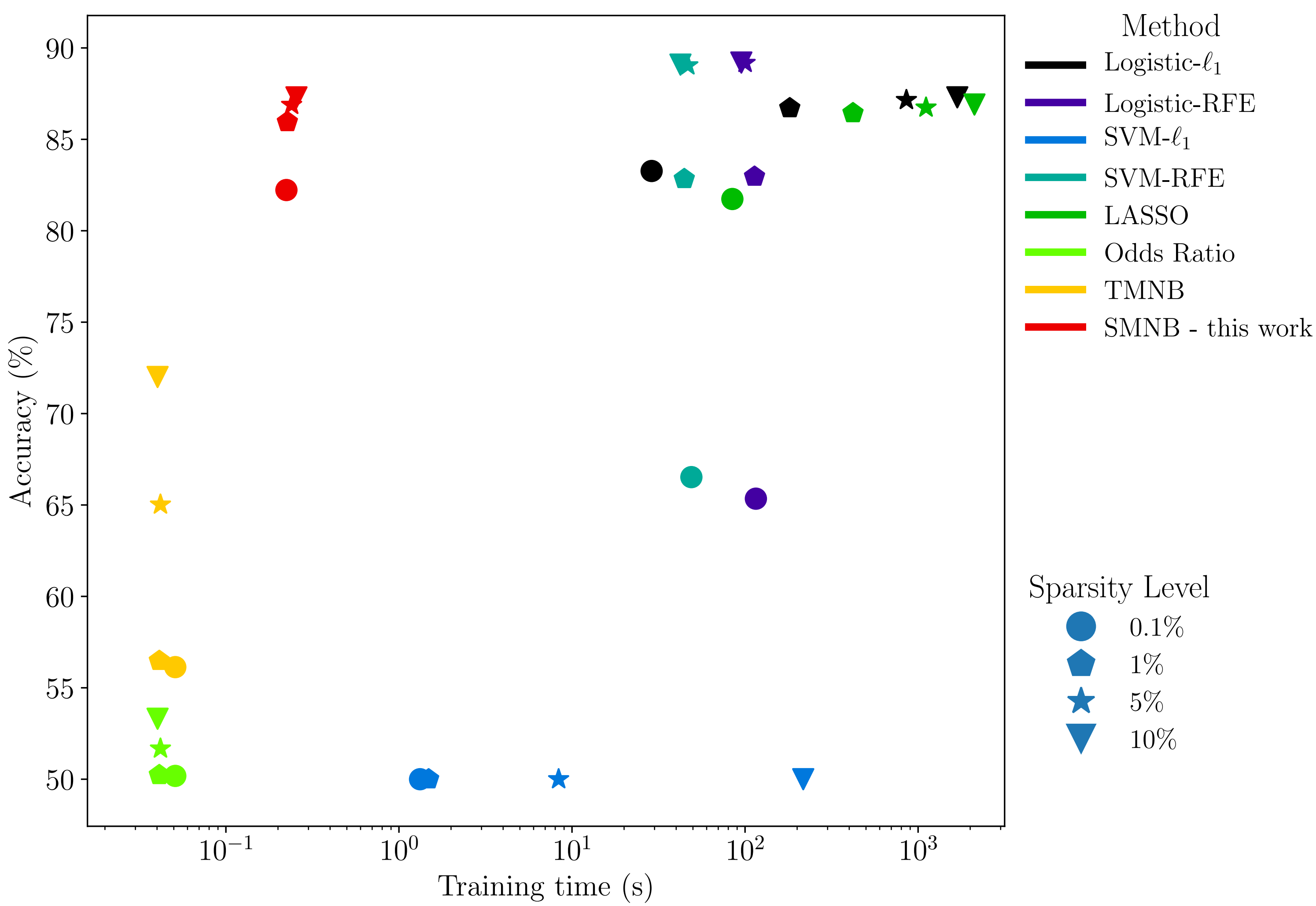
Feature Vectors	Amazon	IMDB	Twitter	MPQA	SST2
Count Vector	31,666	103,124	273,779	6,208	16,599
tf-idf	31,666	103,124	273,779	6,208	16,599
tf-idf wrd bigram	870,536	8,950,169	12,082,555	27,603	227,012
tf-idf char bigram	25,019	48,420	17,812	4838	7762

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

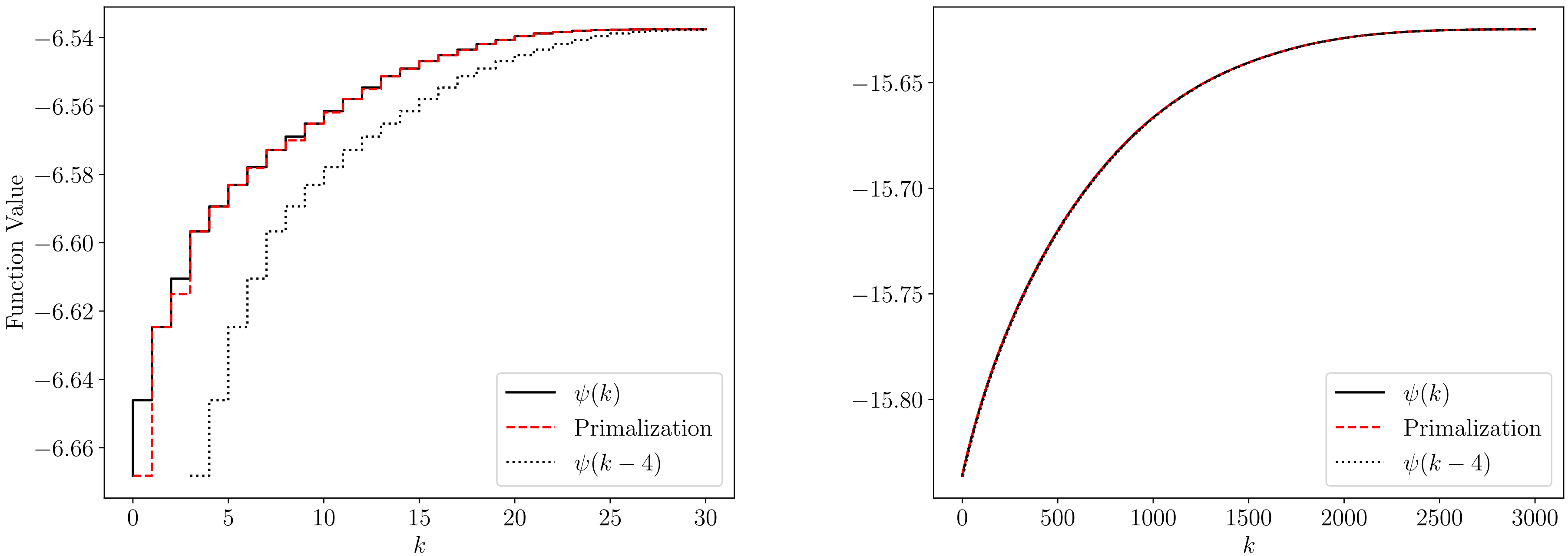
Average run time (seconds, plain Python on CPU).

Naive Feature Selection.



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

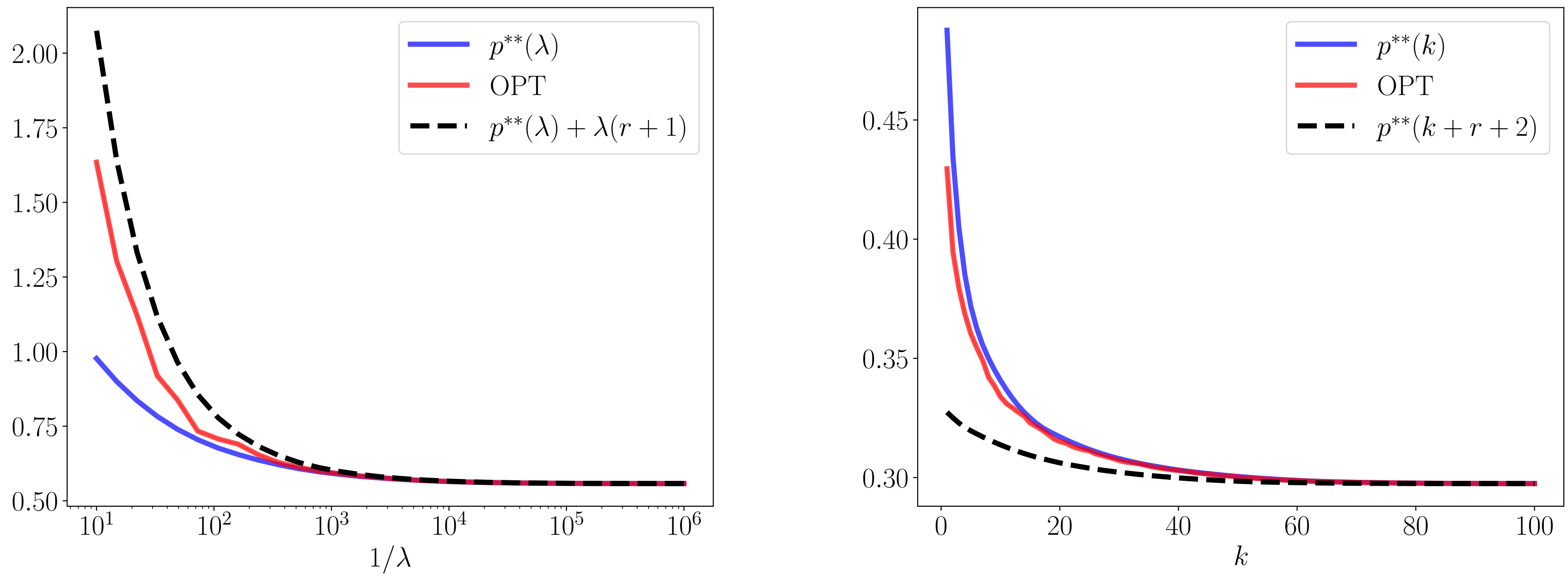
Naive Feature Selection.



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.

LASSO and ℓ_0-Logistic Regression

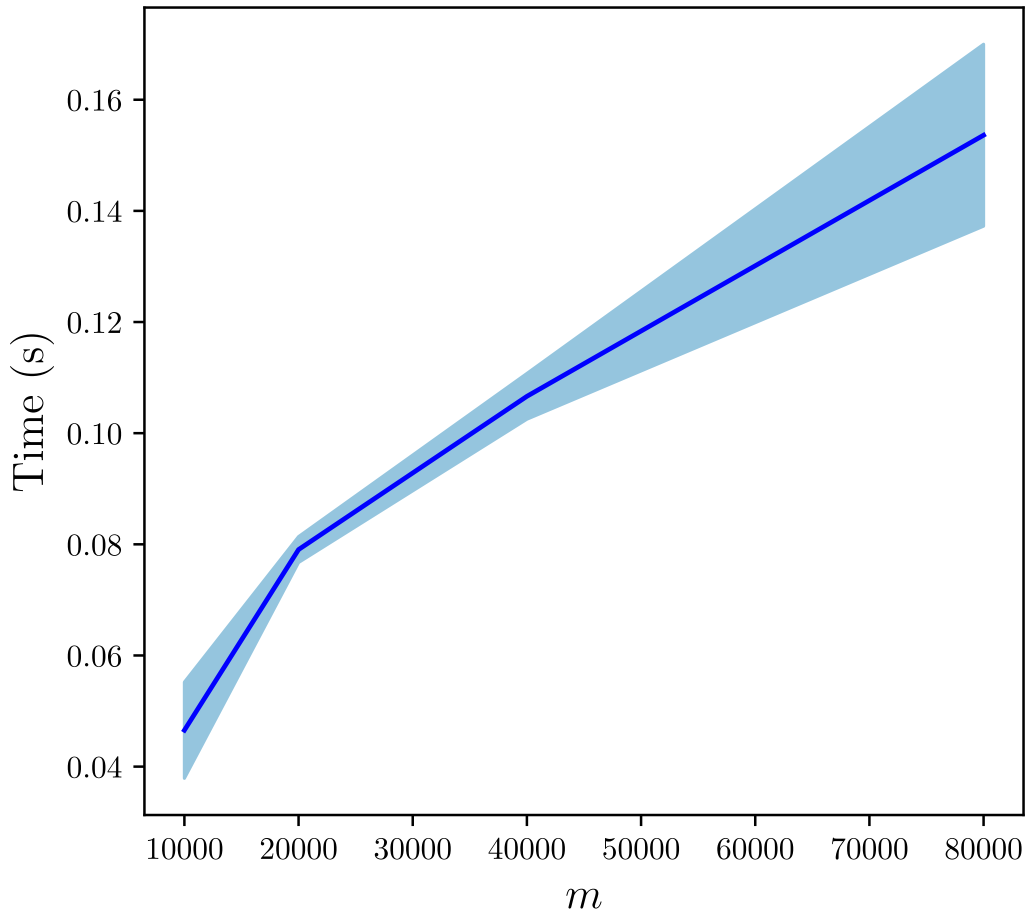
Synthetic example with $X \in \mathbb{R}^{1000 \times 100}$ and rank 10.



Left: Duality gap for linear regression with a ℓ_0 penalty.

Right: Duality gap for ℓ_0 constrained logistic regression.

Naive Feature Selection.

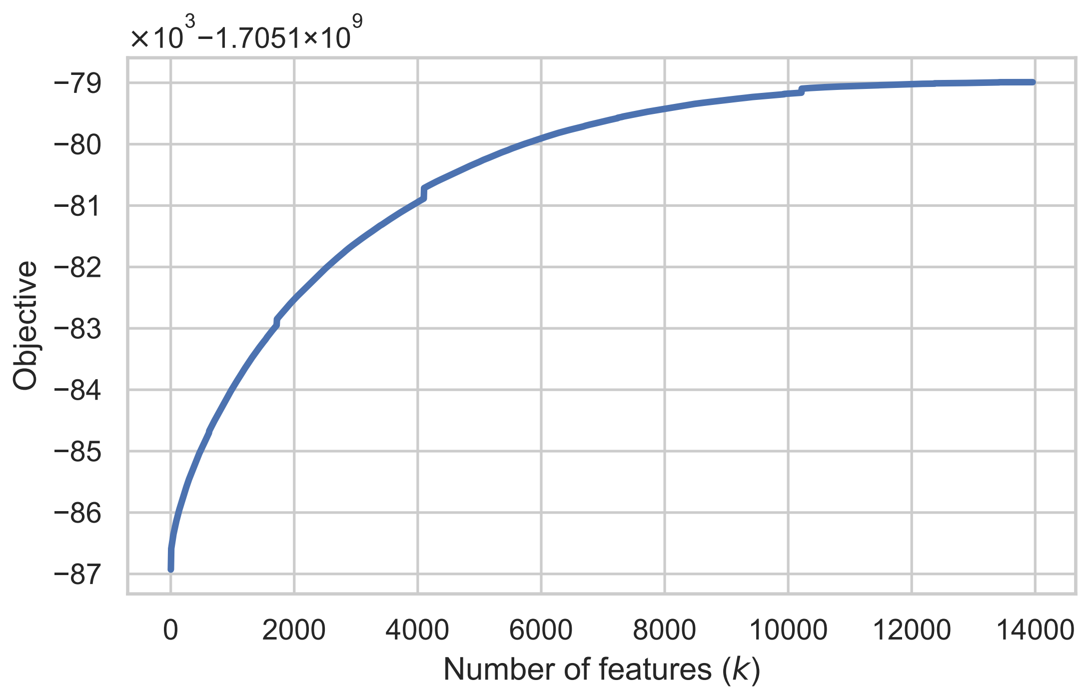


Run time with IMDB dataset/tf-idf vector data set, with increasing m, k with fixed ratio k/m, empirically showing (sub-) linear complexity.

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 ℓ_1 logistic regression, but NFS is $100\times$ faster...
- Requires no RIP assumption (only the naive one behind NB).
- Extends to LASSO, ℓ_0-logistic regression.

Papers: ArXiv:1905.09884. AISTATS 2020 and ArXiv:2102.06742.

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

*

References

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.

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