Structured sparsity-inducing norms through submodular functions

Francis Bach

Willow project, INRIA - Ecole Normale Supérieure

NIPS - December 7, 2010

Outline

- **Introduction: Sparse methods for machine learning**
  - Need for structured sparsity: Going beyond the ℓ_1-norm
- **Submodular functions**
  - Lovász extension
- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Examples
  - Unified algorithms and analysis

Sparsity in supervised machine learning

- Observed data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$
- Regularized empirical risk minimization:

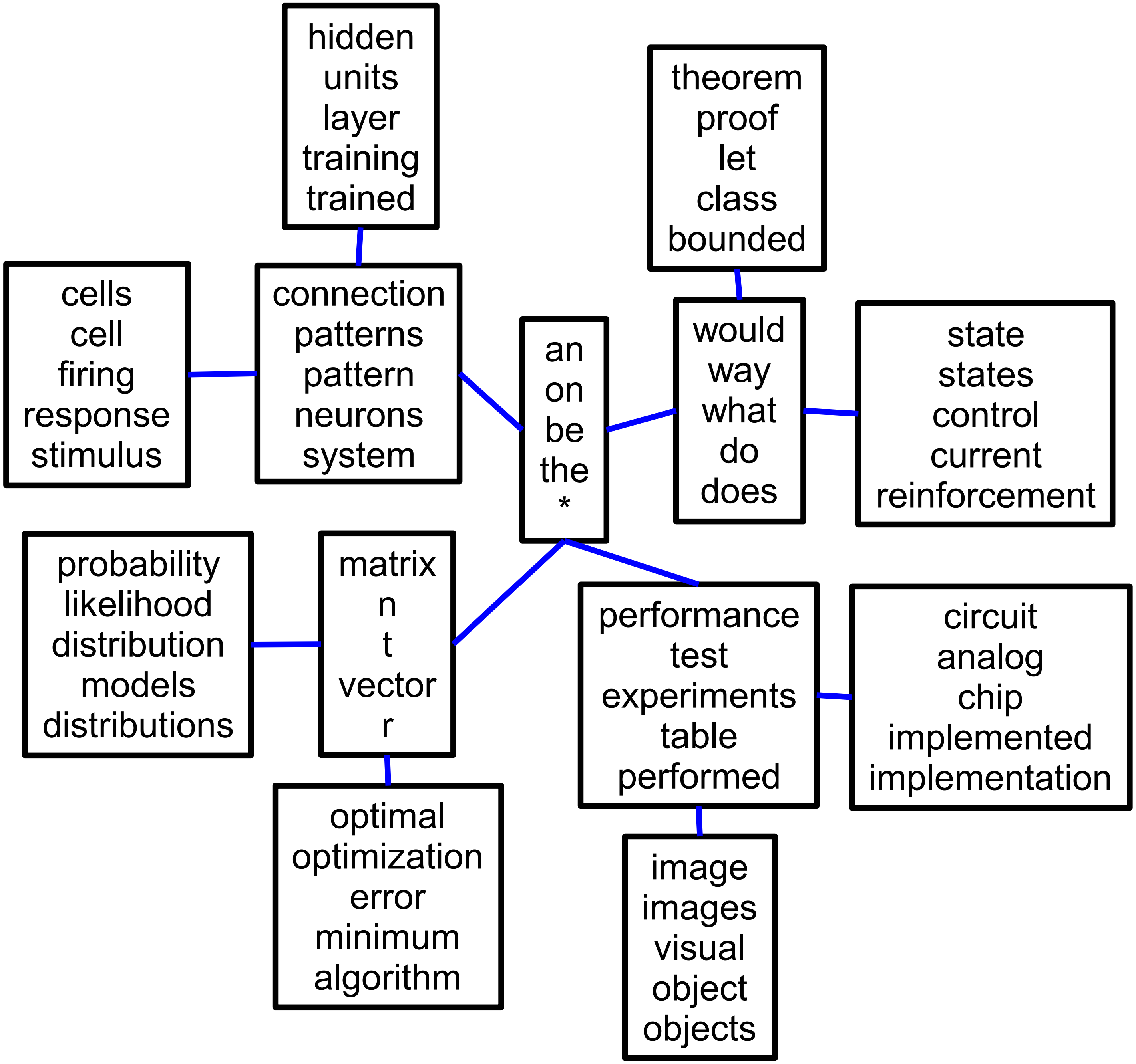
$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \Omega(w)$$

- Norm Ω to promote sparsity
  - square loss + ℓ_1-norm $\Rightarrow$ basis pursuit in signal processing (Chen et al., 2001), Lasso in statistics/machine learning (Tibshirani, 1996)
  - Proxy for interpretability
  - Allow high-dimensional inference: $\boxed{\log p = O(n)}$
- **Generalization to unsupervised learning**
  - dictionary learning/sparse PCA

Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

Modelling of text corpora (Jenatton et al., 2010)



Why structured sparsity?

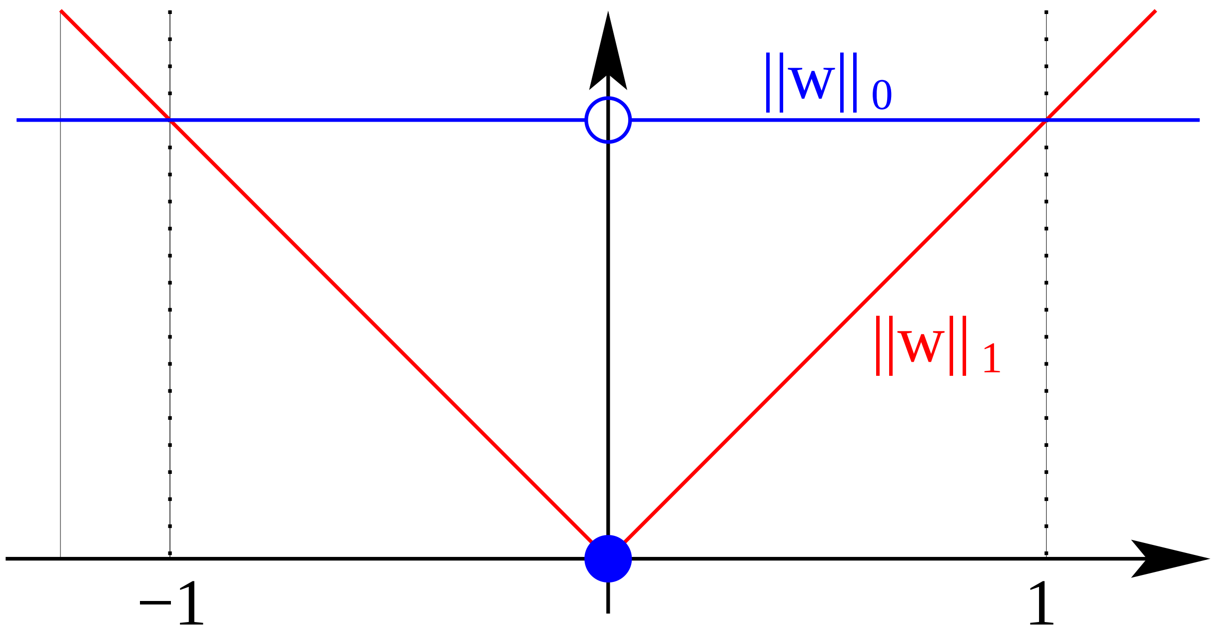
- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

- **Predictive performance**
  - When prior knowledge matches data

- **Numerical efficiency**
  - Non-linear variable selection with 2^p subsets (Bach, 2008)

ℓ_1-norm = convex envelope of cardinality of support

- Let $w \in \mathbb{R}^p$. Let $V = \{1, \dots, p\}$ and $\mathrm{Supp}(w) = \{j \in V,\ w_j \neq 0\}$
- **Cardinality of support**: $\|w\|_0 = \mathrm{Card}(\mathrm{Supp}(w))$
- Convex envelope = largest convex lower bound (see, e.g., Boyd and Vandenberghe, 2004)



- ℓ_1-norm = convex envelope of ℓ_0-quasi-norm on the ℓ_∞-ball $[-1, 1]^p$

Submodular functions (Fujishige, 2005; Bach, 2010b)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\begin{aligned} & \forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B) \\ \Leftrightarrow \ & \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing} \end{aligned}$$

Submodular functions (Fujishige, 2005; Bach, 2010b)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\begin{aligned} & \forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B) \\ \Leftrightarrow \;\; & \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing} \end{aligned}$$

- **Intuition 1**: defined like concave functions (“diminishing returns”)
  - Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave

Submodular functions (Fujishige, 2005; Bach, 2010b)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B)$$
$$\Leftrightarrow \quad \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

- **Intuition 1**: defined like concave functions ("diminishing returns")
  - Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave
- **Intuition 2**: behave like convex functions
  - Polynomial-time minimization, conjugacy theory

Submodular functions (Fujishige, 2005; Bach, 2010b)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B)$$
$$\Leftrightarrow \quad \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

- **Intuition 1**: defined like concave functions ("diminishing returns")
  - Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave

- **Intuition 2**: behave like convex functions
  - Polynomial-time minimization, conjugacy theory

- Used in several areas of signal processing and machine learning
  - Total variation/graph cuts (Chambolle, 2005; Boykov et al., 2001)
  - Optimal design (Krause and Guestrin, 2005)

Submodular functions - Lovász extension

- Given any set-function F and w such that $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$, define:

$$f(w) = \sum_{k=1}^{p} w_{j_k} [F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})]$$

  - If $w = 1_A$, $f(w) = F(A) \Rightarrow$ extension from $\{0,1\}^p$ to $\mathbb{R}^p$
  - f is piecewise affine and positively homogeneous

- F is submodular if and only if f is convex
  - Minimizing $f(w)$ on $w \in [0,1]^p$ equivalent to minimizing F on 2^V

Submodular functions and structured sparsity

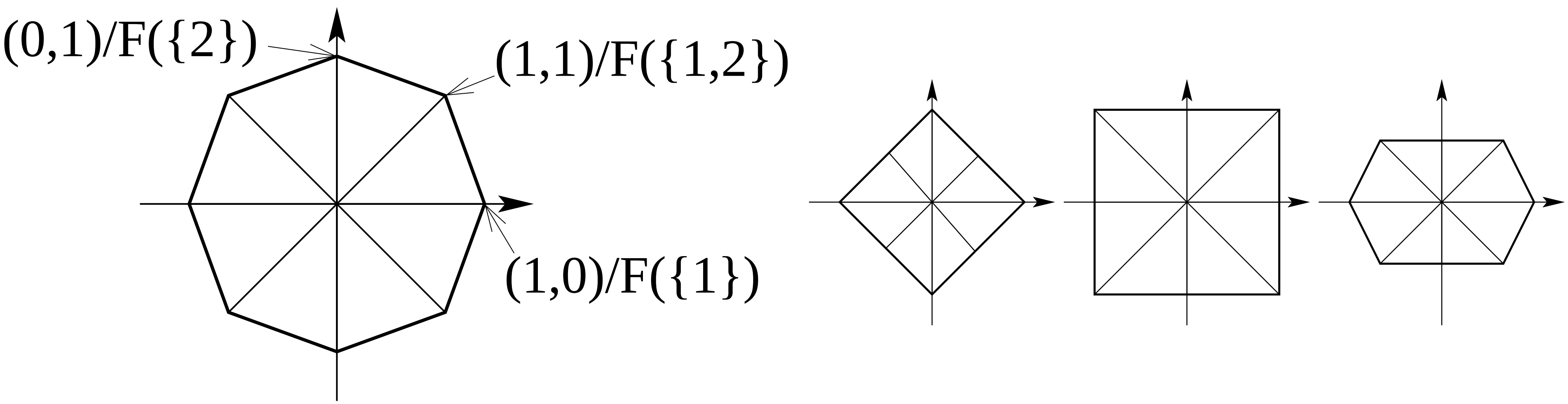
- Let $F : 2^V \to \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\mathrm{Supp}(w))$ on the ℓ_∞-ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F

Submodular functions and structured sparsity

- Let $F : 2^V \to \mathbb{R}$ be a **non-decreasing submodular set-function**

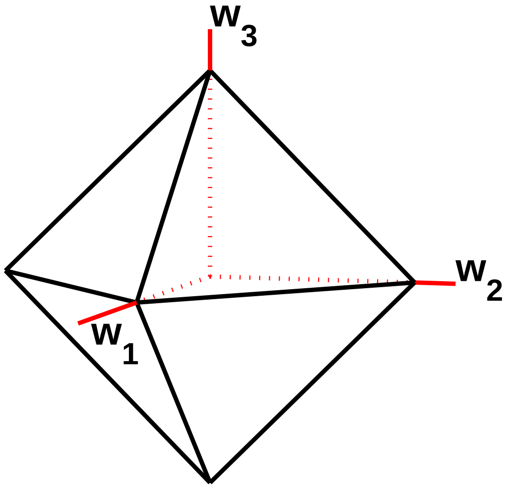
- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\mathrm{Supp}(w))$ on the ℓ_∞-ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F

- **Sparsity-inducing properties**: Ω is a polyhedral norm

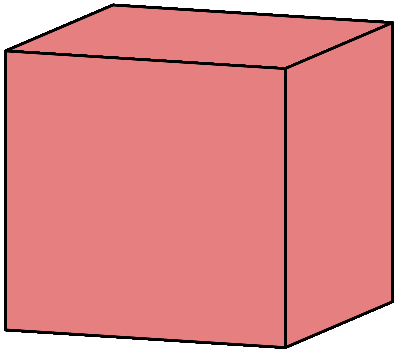


  – A if stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
  – With probability one, stable sets are the only allowed active sets

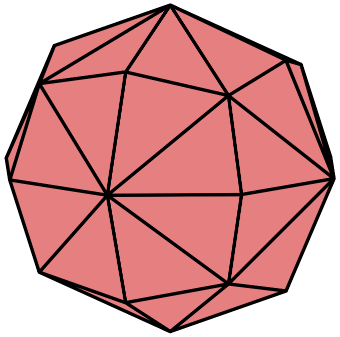
Polyhedral unit balls



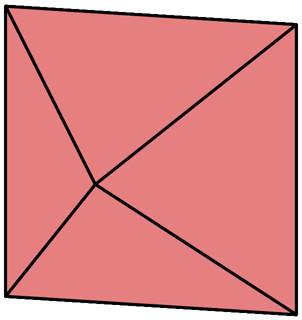
$F(A) = |A|$
$\Omega(w) = \|w\|_1$



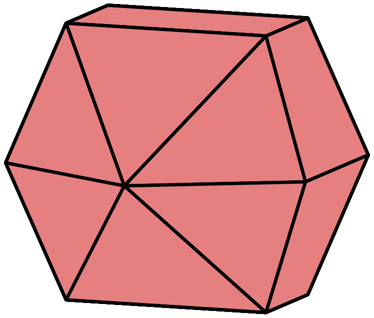
$F(A) = \min\{|A|, 1\}$
$\Omega(w) = \|w\|_\infty$



$F(A) = |A|^{1/2}$
all possible extreme points



$F(A) = 1_{\{A\cap\{1\}\neq\varnothing\}} + 1_{\{A\cap\{2,3\}\neq\varnothing\}}$
$\Omega(w) = |w_1| + \|w_{\{2,3\}}\|_\infty$



$F(A) = 1_{\{A\cap\{1,2,3\}\neq\varnothing\}}$
$+1_{\{A\cap\{2,3\}\neq\varnothing\}} + 1_{\{A\cap\{3\}\neq\varnothing\}}$
$\Omega(w) = \|w\|_\infty + \|w_{\{2,3\}}\|_\infty + |w_3|$

Submodular functions and structured sparsity
Examples

- **From $\Omega(w)$ to $F(A)$**: provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathcal{G}} \|w_G\|_\infty$$

  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero: $\mathrm{Supp}(w)^c = \bigcup_{G \in \mathcal{H}} G$ for some $\mathcal{H} \subseteq \mathcal{G}$

Submodular functions and structured sparsity
Examples

- **From $\Omega(w)$ to $F(A)$**: provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathcal{G}} \|w_G\|_\infty$$

  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero: $\mathrm{Supp}(w)^c = \bigcup_{G \in \mathcal{H}} G$ for some $\mathcal{H} \subseteq \mathcal{G}$

- Associated submodular function

$$F(A) = \mathrm{Card}\big(\{G \in \mathcal{G},\ G \cap A \neq \varnothing\}\big)$$

  - **Justification not only limited to allowed sparsity patterns**

Submodular functions and structured sparsity
Examples

- **From $\Omega(w)$ to $F(A)$**: provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathcal{G}} \|w_G\|_\infty \quad \Rightarrow \quad F(A) = \mathrm{Card}\big(\{G \in \mathcal{G},\ G \cap A \neq \varnothing\}\big)$$

Submodular functions and structured sparsity
Examples

- **From $\Omega(w)$ to $F(A)$:** provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathcal{G}} \|w_G\|_\infty \quad \Rightarrow \quad F(A) = \mathrm{Card}\big(\{G \in \mathcal{G},\ G \cap A \neq \varnothing\}\big)$$

- **From $F(A)$ to $\Omega(w)$:** provides new sparsity-inducing norms
  - $F(A) = g(\mathrm{Card}(A)) \ \Rightarrow\ \Omega$ is a combination of **order statistics**
  - **Non-factorial priors** for supervised learning: Ω depends on the eigenvalues of $X_A^\top X_A$ and not simply on the cardinality of A

Non-factorial priors for supervised learning

- Selection of subset A from design matrix $X \in \mathbb{R}^{n \times p}$
- **Frequentist analysis** (Mallow's C_L): $\operatorname{tr} X_A^\top X_A (X_A^\top X_A + \lambda I)^{-1}$
  - Not submodular
- **Bayesian analysis** (marginal likelihood): $\log\det(X_A^\top X_A + \lambda I)$
  - Submodular (also true for $\operatorname{tr}(X_A^\top X_A)^{1/2}$)

p	n	k	submod.	ℓ_2 vs. submod.	ℓ_1 vs. submod.	greedy vs. submod.
120	120	80	40.8 ± 0.8	-2.6 ± 0.5	**0.6** ± **0.0**	**21.8** ± **0.9**
120	120	40	35.9 ± 0.8	**2.4** ± **0.4**	**0.3** ± **0.0**	**15.8** ± **1.0**
120	120	20	29.0 ± 1.0	**9.4** ± **0.5**	-0.1 ± 0.0	**6.7** ± **0.9**
120	120	10	20.4 ± 1.0	**17.5** ± **0.5**	-0.2 ± 0.0	-2.8 ± 0.8
120	20	20	49.4 ± 2.0	0.4 ± 0.5	**2.2** ± **0.8**	**23.5** ± **2.1**
120	20	10	49.2 ± 2.0	0.0 ± 0.6	1.0 ± 0.8	**20.3** ± **2.6**
120	20	6	43.5 ± 2.0	**3.5** ± **0.8**	**0.9** ± **0.6**	**24.4** ± **3.0**
120	20	4	41.0 ± 2.1	**4.8** ± **0.7**	-1.3 ± 0.5	**25.1** ± **3.5**

Unified optimization algorithms

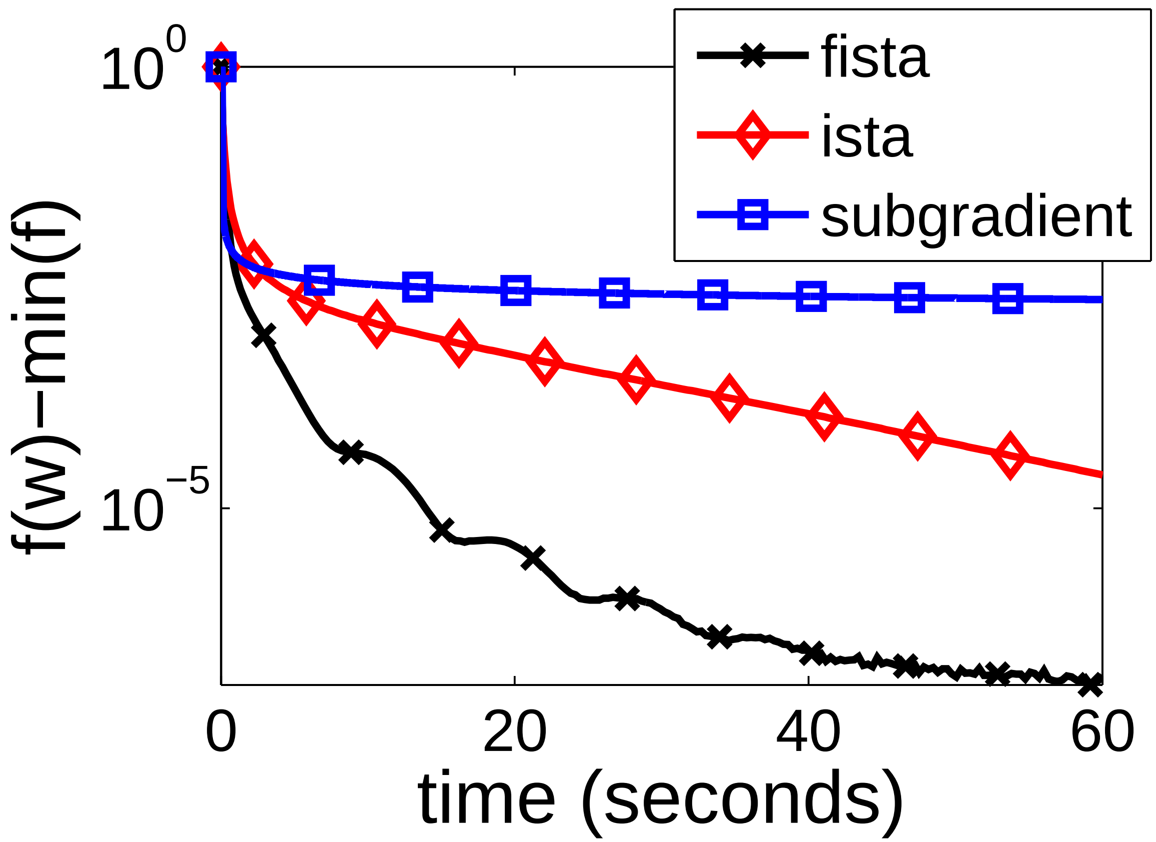
- **Polyhedral norm** with $O(3^p)$ extreme points
  - Not suitable to linear programming toolboxes
- **Subgradient** ($w \mapsto \Omega(w)$ non-differentiable)
  - subgradient may be obtained in polynomial time $\Rightarrow$ too slow

Unified optimization algorithms

- **Polyhedral norm** with $O(3^p)$ extreme points
  - Not suitable to linear programming toolboxes
- **Subgradient** ($w \mapsto \Omega(w)$ non-differentiable)
  - subgradient may be obtained in polynomial time $\Rightarrow$ too slow
- **Proximal methods** (e.g., Beck and Teboulle, 2009)
  - $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda\Omega(w)$: differentiable + non-differentiable
  - Efficient when (P) : $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - v\|_2^2 + \lambda\Omega(w)$ is "easy"
- **Proposition**: (P) is equivalent to $\min_{A \subset V} \lambda F(A) - \sum_{j \in A} |v_j|$ with minimum-norm-point algorithm
  - No complexity bounds, but empirically $O(p^2)$
  - Faster algorithm for special case: poster T24 (Mairal et al., 2010)

Comparison of optimization algorithms

- Synthetic example with $p = 1000$ and $F(A) = |A|^{1/2}$
- ISTA: proximal method
- FISTA: accelerated variant (Beck and Teboulle, 2009)



Unified theoretical analysis

- **Decomposability**
  - Key to theoretical analysis (Negahban et al., 2009)
  - **Property**: $\forall w \in \mathbb{R}^p$, and $\forall J \subset V$, if $\min_{j \in J} |w_j| \geqslant \max_{j \in J^c} |w_j|$, then $\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c})$

- **Support recovery**
  - Extension of known sufficient condition (Zhao and Yu, 2006; Negahban and Wainwright, 2008)

- **High-dimensional inference**
  - Extension of known sufficient condition (Bickel et al., 2009)
  - Matches with analysis of Negahban et al. (2009) for common cases

Conclusion

- **Structured sparsity through submodular functions**
  - Many applications (image, audio, text, etc.)
  - Unified analysis and algorithms

Conclusion

- **Structured sparsity through submodular functions**
  - Many applications (image, audio, text, etc.)
  - Unified analysis and algorithms

- **On-going work on structured sparsity**
  - Extension to symmetric submodular functions (Bach, 2010a)
    * Shaping all level sets $\{w = \alpha\}$, $\alpha \in \mathbb{R}$, rather than only $\alpha = 0$
  - Norm design beyond submodular functions
  - Links with greedy methods (Haupt and Nowak, 2006; Huang et al., 2009)
  - Extensions to matrices

References

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008.

F. Bach. Shaping level sets with submodular functions. Technical Report 00542949, HAL, 2010a.

F. Bach. Convex analysis and optimization with submodular functions: a tutorial. Technical Report 00527714, HAL, 2010b.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.

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.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

R. Jenatton, J.Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, arXiv:0904.3523, 2009a.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical report, arXiv:0909.1440, 2009b.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Submitted to ICML*, 2010.

K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *Proceedings of CVPR*, 2009.

A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proc. UAI*, 2005.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *NIPS*, 2010.

S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of ℓ_1-ℓ_∞-regularization. In *Adv. NIPS*, 2008.

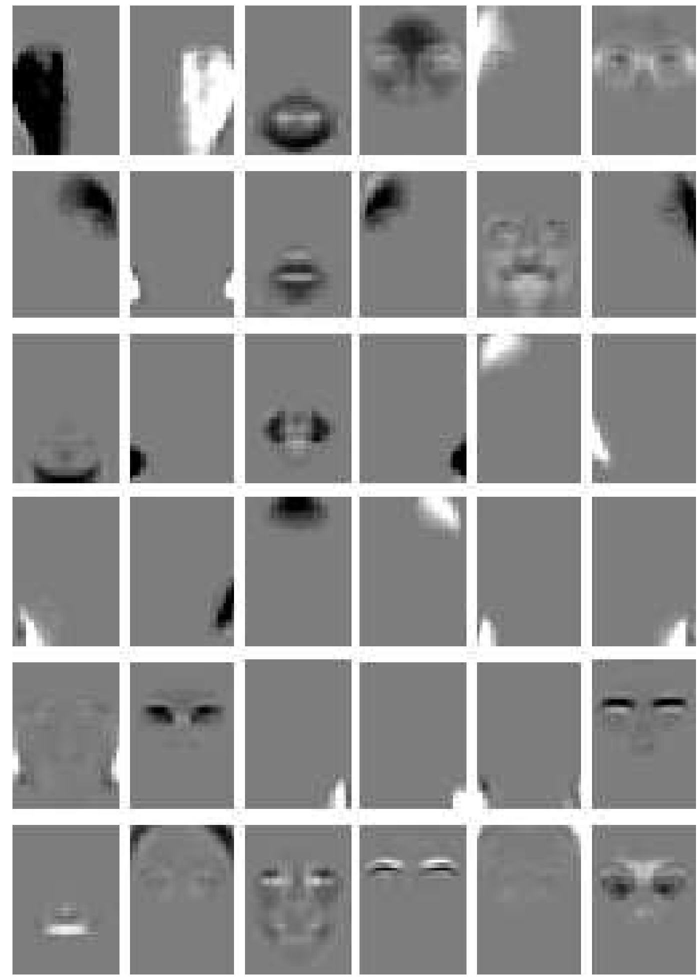
S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Structured sparse PCA (Jenatton et al., 2009b)

raw data

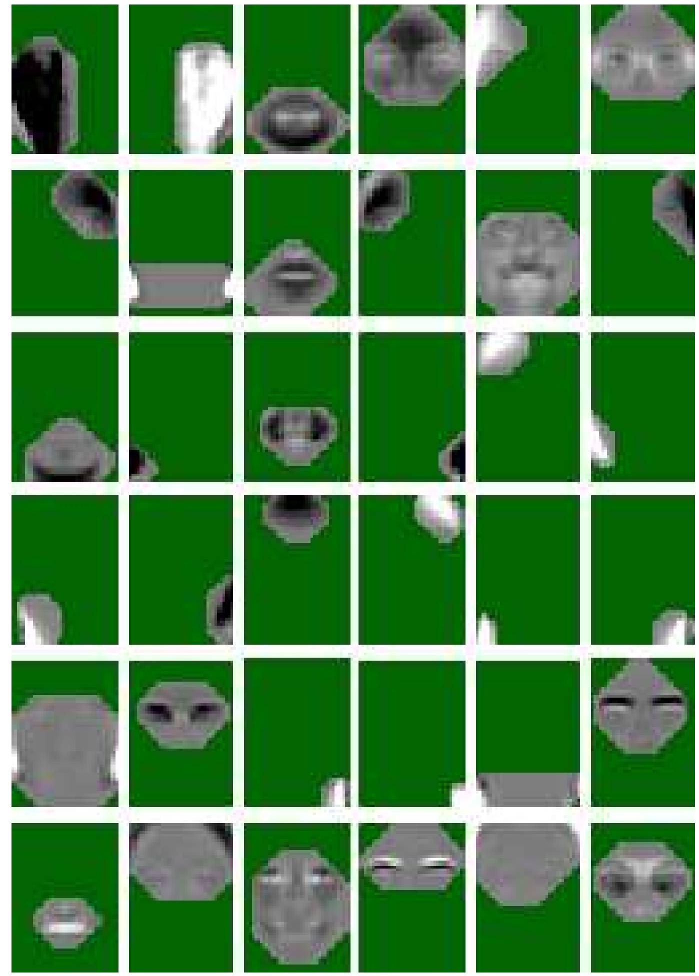


Structured sparse PCA

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion in face identification

Structured sparse PCA (Jenatton et al., 2009b)

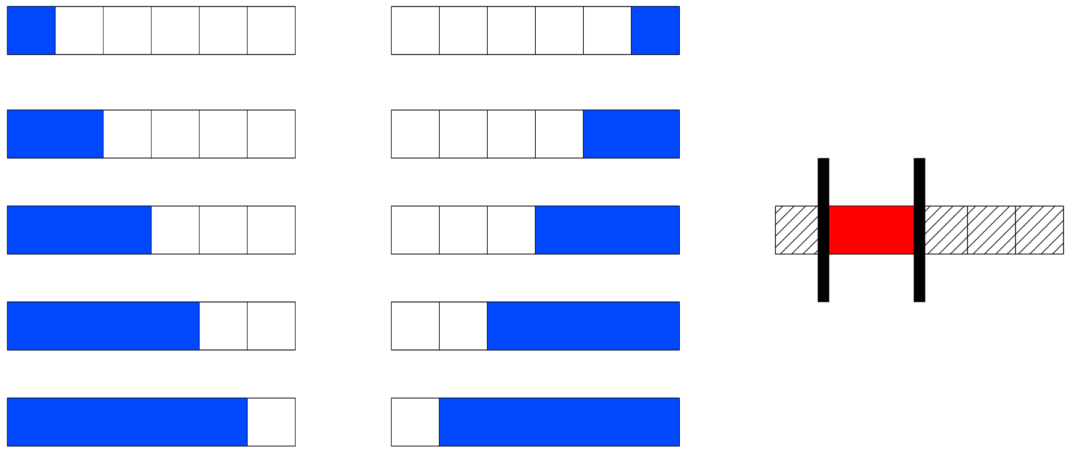
raw data



Structured sparse PCA

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion in face identification

Selection of contiguous patterns in a sequence

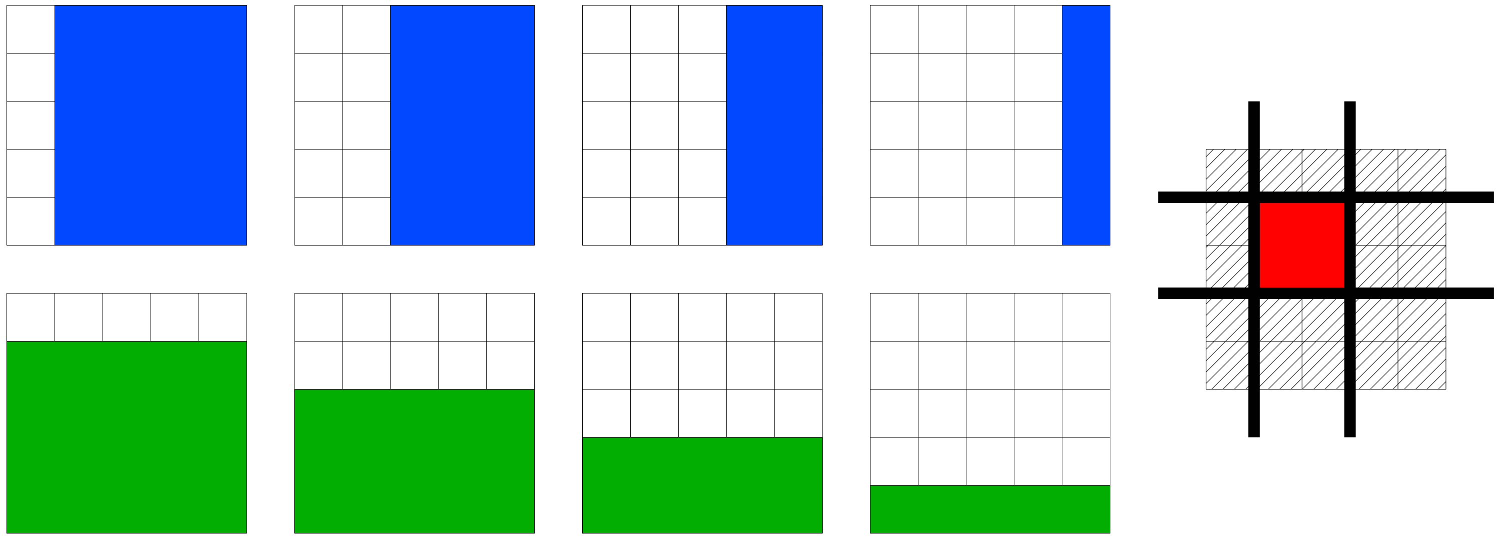


- $\mathcal{G}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a **contiguous pattern**

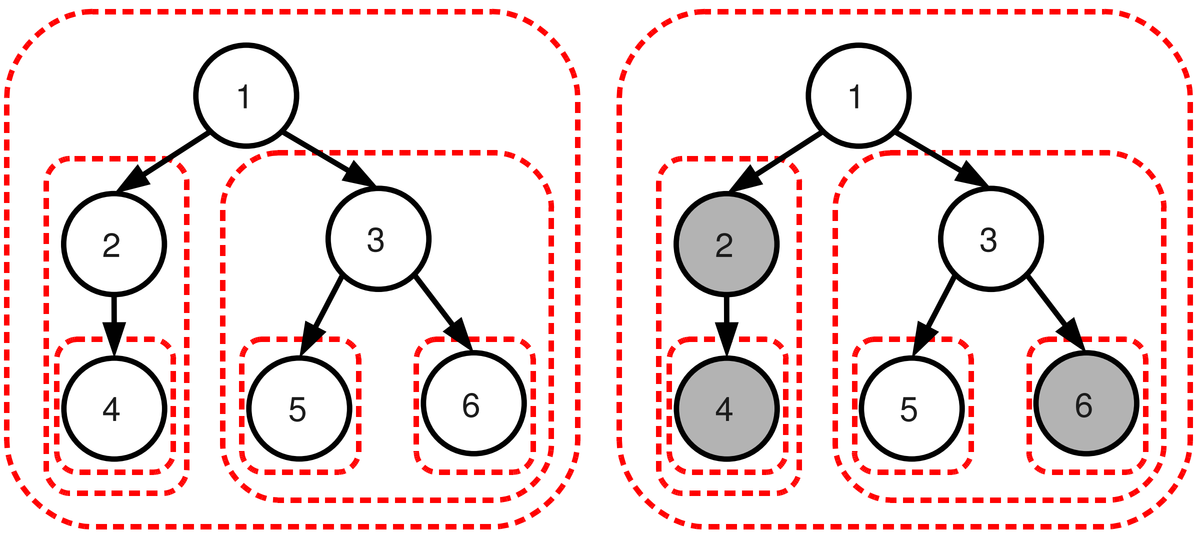
- $\sum_{G\in\mathcal{G}} \|w_G\|_\infty \;\Rightarrow\; F(A) = p - 2 + \text{Range}(A)$ if $A \neq \varnothing$, $F(\varnothing) = 0$
  - Jump from 0 to $p-1$: tends to include all variables simultaneously
  - Add $\nu|A|$ to smooth the kink: all sparsity patterns are possible
  - **Contiguous patterns are favored (and not forced)**

Extensions of norms with overlapping groups

- Selection of **rectangles** (at any position) in a 2-D grids



- **Hierarchies**



Support recovery - $\min_{w\in\mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w)$

- **Notation**
  - $\rho(J) = \min_{B\subset J^c} \frac{F(B\cup J)-F(J)}{F(B)} \in (0,1]$ (for J stable)
  - $c(J) = \sup_{w\in\mathbb{R}^p} \Omega_J(w_J)/\|w_J\|_2 \leqslant |J|^{1/2} \max_{k\in V} F(\{k\})$

- **Proposition**
  - Assume $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0,I)$
  - J = smallest stable set containing the support of w^*
  - Assume $\nu = \min_{j,w_j^*\neq 0} |w_j^*| > 0$
  - Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{p\times p}$. Assume $\kappa = \lambda_{\min}(Q_{JJ}) > 0$
  - Assume that for $\eta > 0$, $\boxed{(\Omega^J)^*[(\Omega_J(Q_{JJ}^{-1}Q_{Jj}))_{j\in J^c}] \leqslant 1-\eta}$
  - If $\lambda \leqslant \frac{\kappa\nu}{2c(J)}$, $\hat{w}$ has support equal to J, with probability larger than
$$1 - 3P\Big(\Omega^*(z) > \frac{\lambda\eta\rho(J)\sqrt{n}}{2\sigma}\Big)$$
  - z is a multivariate normal with covariance matrix Q

Consistency - $\min_{w\in\mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w)$

- **Proposition**
  - Assume $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I)$
  - J = smallest stable set containing the support of w^*
  - Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{p\times p}$.
  - Assume that $\forall\Delta$ s.t. $\Omega^J(\Delta_{J^c}) \leqslant 3\Omega_J(\Delta_J)$, $\Delta^\top Q\Delta \geqslant \kappa\|\Delta_J\|_2^2$
  - Then $\boxed{\Omega(\hat{w} - w^*) \leqslant \dfrac{24c(J)^2\lambda}{\kappa\rho(J)^2}}$ and $\boxed{\dfrac{1}{n}\|X\hat{w} - Xw^*\|_2^2 \leqslant \dfrac{36c(J)^2\lambda^2}{\kappa\rho(J)^2}}$

    with probability larger than $1 - P\big(\Omega^*(z) > \frac{\lambda\rho(J)\sqrt{n}}{2\sigma}\big)$
  - z is a multivariate normal with covariance matrix Q

- **Concentration inequality** (z normal with covariance matrix Q):
  - $\mathcal{T}$ set of stable inseparable sets
  - Then $P(\Omega^*(z) > t) \leqslant \sum_{A\in\mathcal{T}} 2^{|A|} \exp\big(-\frac{t^2F(A)^2/2}{1^\top Q_{AA}1}\big)$