Sparse Estimation and Dictionary Learning

(for Biostatistics?)

Julien Mairal

Biostatistics Seminar, UC Berkeley

What this talk is about?

- **Why sparsity, what for and how?**
- **Feature learning / clustering / sparse PCA**;
- **Machine learning**: selecting relevant features;
- **Signal and image processing**: restoration, reconstruction;
- **Biostatistics**: you tell me.

Part I: Sparse Estimation

Sparse Linear Model: Machine Learning Point of View

Let $(y^i, \mathbf{x}^i)_{i=1}^n$ be a training set, where the vectors $\mathbf{x}^i$ are in $\mathbb{R}^p$ and are called features. The scalars y^i are in

- $\{-1, +1\}$ for **binary** classification problems.
- $\mathbb{R}$ for **regression** problems.

We assume there is a relation $y \approx \mathbf{w}^\top \mathbf{x}$, and solve

$$\min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y^i, \mathbf{w}^\top \mathbf{x}^i)}_{\text{empirical risk}} + \underbrace{\lambda \psi(\mathbf{w})}_{\text{regularization}} .$$

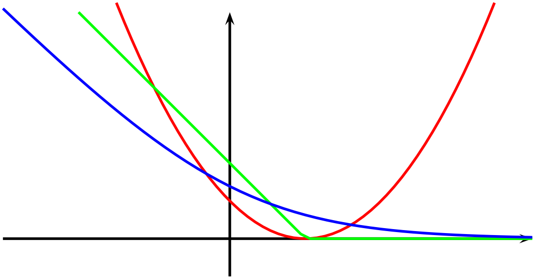
Sparse Linear Models: Machine Learning Point of View

A few examples:

Ridge regression:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y^i - \mathbf{w}^\top \mathbf{x}^i)^2 + \lambda\|\mathbf{w}\|_2^2.$$

Linear SVM:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y^i \mathbf{w}^\top \mathbf{x}^i) + \lambda\|\mathbf{w}\|_2^2.$$

Logistic regression:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y^i \mathbf{w}^\top \mathbf{x}^i}\right) + \lambda\|\mathbf{w}\|_2^2.$$



Sparse Linear Models: Machine Learning Point of View

A few examples:

Ridge regression:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y^i - \mathbf{w}^\top \mathbf{x}^i)^2 + \lambda\|\mathbf{w}\|_2^2.$$

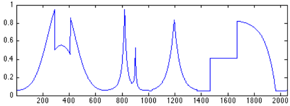
Linear SVM:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y^i\mathbf{w}^\top \mathbf{x}^i) + \lambda\|\mathbf{w}\|_2^2.$$

Logistic regression:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y^i\mathbf{w}^\top \mathbf{x}^i}\right) + \lambda\|\mathbf{w}\|_2^2.$$

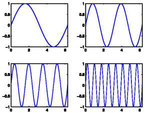
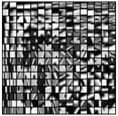
The **squared ℓ_2-norm** induces "**smoothness**" in $\mathbf{w}$. When one knows in advance that $\mathbf{w}$ should be sparse, one should use a **sparsity-inducing** regularization such as the **ℓ_1-norm**. [Chen et al., 1999, Tibshirani, 1996]

Sparse Linear Models: Signal Processing Point of View

Let $\mathbf{y}$ in $\mathbb{R}^n$ be a signal.




Let $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^p] \in \mathbb{R}^{n \times p}$ be a set of normalized "basis vectors".
We call it **dictionary**.

$\mathbf{X}$ is "adapted" to $\mathbf{y}$ if it can represent it with a few basis vectors—that is, there exists a **sparse vector** $\mathbf{w}$ in $\mathbb{R}^p$ such that $\mathbf{y} \approx \mathbf{X}\mathbf{w}$. We call $\mathbf{w}$ the **sparse code**.

$$\underbrace{\begin{pmatrix} \mathbf{y} \end{pmatrix}}_{\mathbf{y} \in \mathbb{R}^n} \approx \underbrace{\left(\begin{array}{c|c|c|c} \mathbf{x}^1 & \mathbf{x}^2 & \cdots & \mathbf{x}^p \end{array}\right)}_{\mathbf{X} \in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_p \end{pmatrix}}_{\mathbf{w} \in \mathbb{R}^p, \textbf{sparse}}$$

The Sparse Decomposition Problem

$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2}_{\text{data fitting term}} + \underbrace{\lambda\psi(\mathbf{w})}_{\text{sparsity-inducing regularization}}$$

ψ induces sparsity in $\mathbf{w}$. It can be

- the ℓ_0 "pseudo-norm". $\|\mathbf{w}\|_0 \triangleq \#\{i \text{ s.t. } \mathbf{w}_i \neq 0\}$ (NP-hard)
- the ℓ_1 norm. $\|\mathbf{w}\|_1 \triangleq \sum_{i=1}^p |\mathbf{w}_i|$ (convex),
- ...

This is a selection problem. When ψ is the ℓ_1-norm, the problem is called **Lasso** [Tibshirani, 1996] or **basis pursuit** [Chen et al., 1999]

Why does the ℓ_1-norm induce sparsity?

Exemple: quadratic problem in 1D

$$\min_{w \in \mathbb{R}} \frac{1}{2}(u-w)^2 + \lambda |w|$$

Piecewise quadratic function with a kink at zero.

Derivative at 0_+: $g_+ = -u + \lambda$ and 0_-: $g_- = -u - \lambda$.

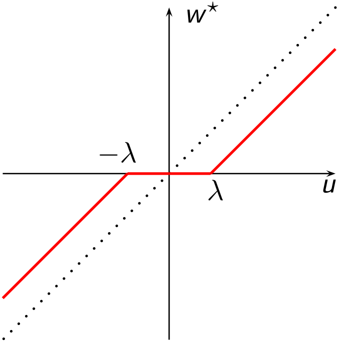
Optimality conditions. w is optimal iff:

- $|w| > 0$ and $(u-w) + \lambda \operatorname{sign}(w) = 0$
- $w = 0$ and $g_+ \geq 0$ and $g_- \leq 0$

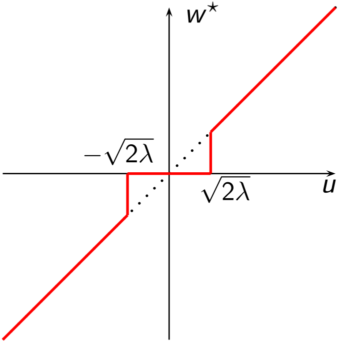
The solution is the **soft-thresholding operator**:

$$w^\star = \operatorname{sign}(u)(|u| - \lambda)^+.$$

Why does the ℓ_1-norm induce sparsity?



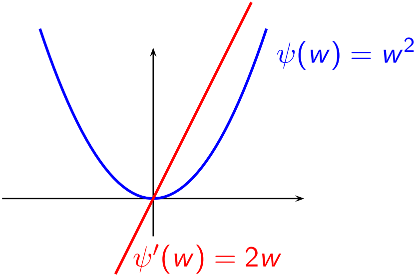
(a) soft-thresholding operator,
$w^\star = \text{sign}(u)(|u| - \lambda)^+$,
$\min_w \frac{1}{2}(u-w)^2 + \lambda|w|$

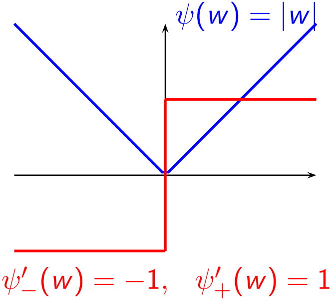


(b) hard-thresholding operator
$w^\star = \mathbf{1}_{|u| \geq \sqrt{2\lambda}} u$
$\min_w \frac{1}{2}(u-w)^2 + \lambda \mathbf{1}_{|w|>0}$

Why does the ℓ_1-norm induce sparsity?

Comparison with ℓ_2-regularization in 1D

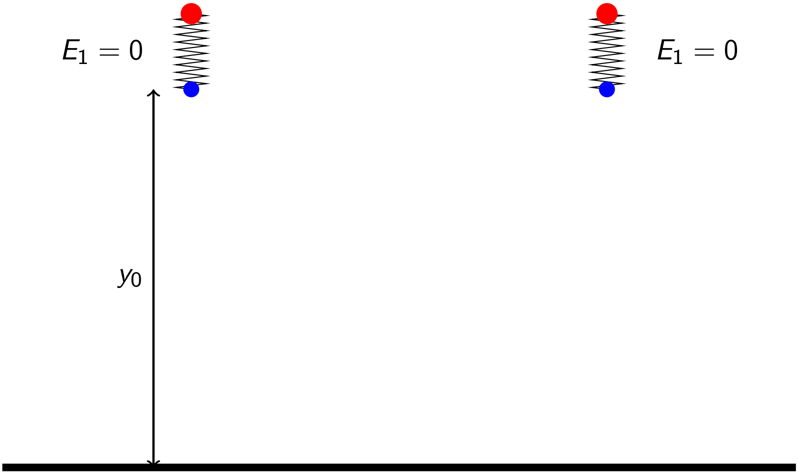




The gradient of the ℓ_2-penalty vanishes when w get close to 0. On its differentiable part, the norm of the gradient of the ℓ_1-norm is constant.

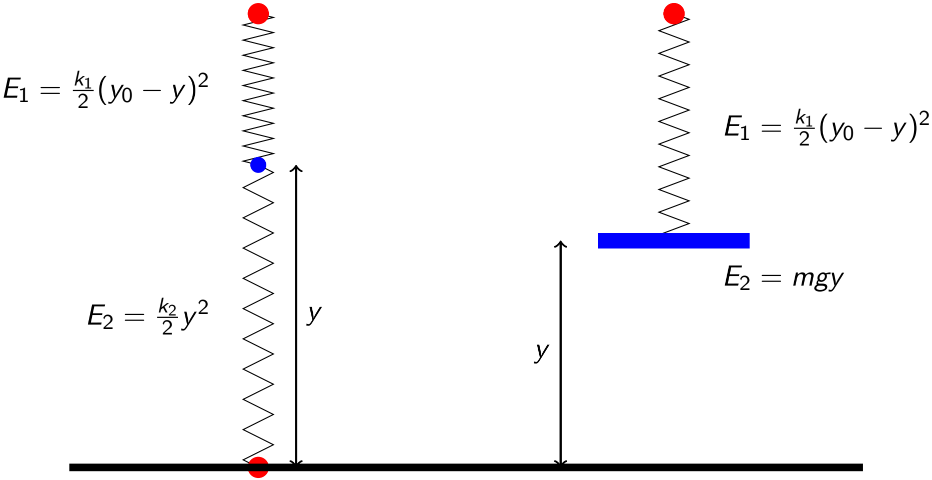
Why does the ℓ_1-norm induce sparsity?

Physical illustration



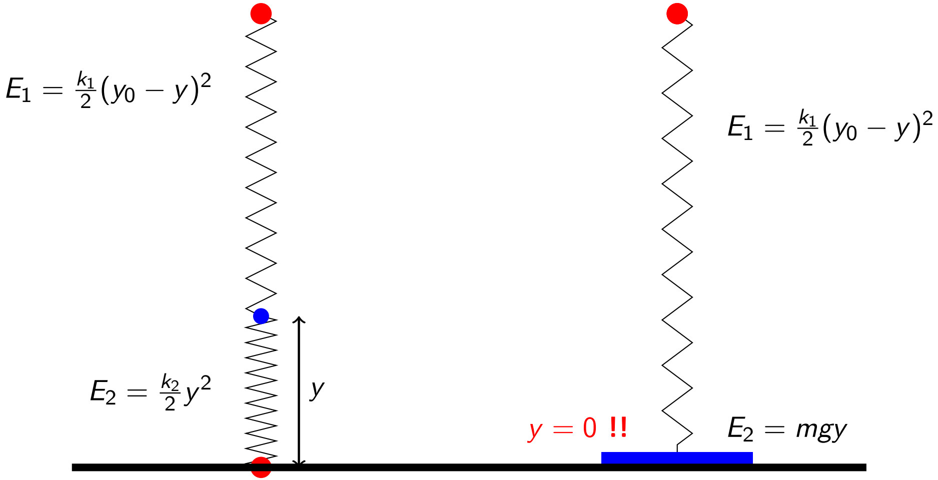
Why does the ℓ_1-norm induce sparsity?

Physical illustration

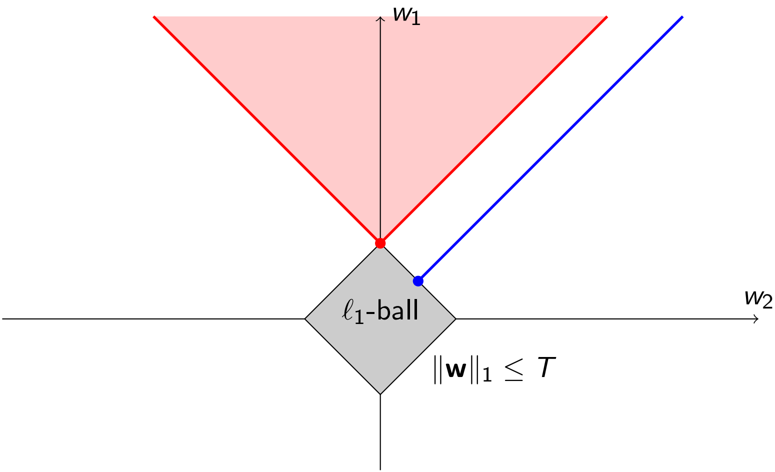


Why does the ℓ_1-norm induce sparsity?

Physical illustration

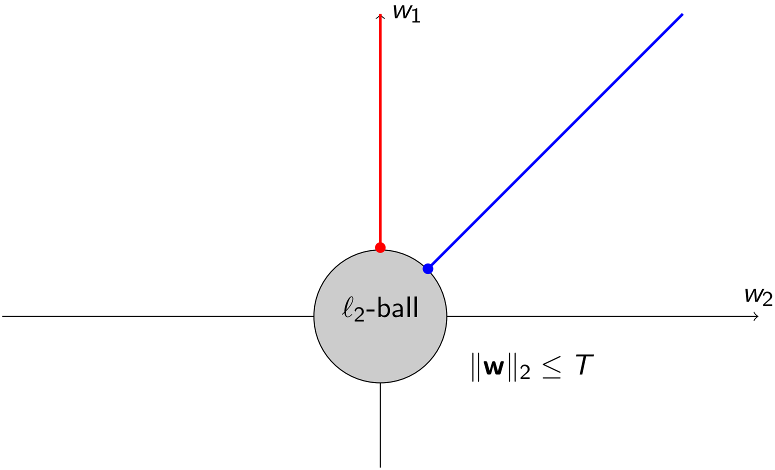


Regularizing with the ℓ_1-norm



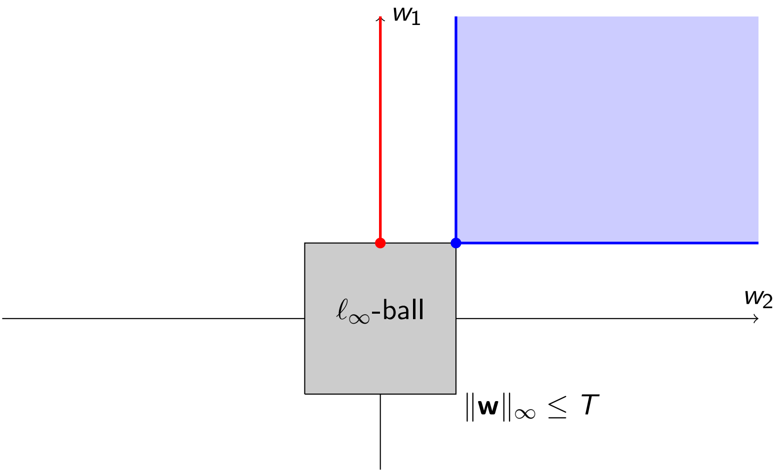
The projection onto a convex set is "biased" towards singularities.

Regularizing with the ℓ_2-norm



The ℓ_2-norm is isotropic.

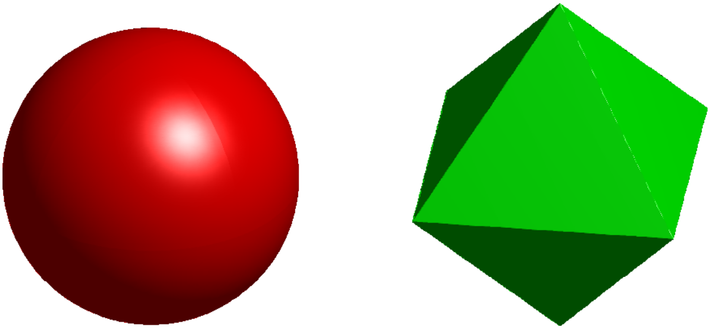
Regularizing with the ℓ_∞-norm



The ℓ_∞-norm encourages $|w_1| = |w_2|$.

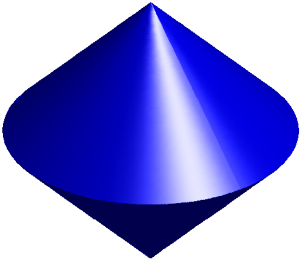
In 3D.

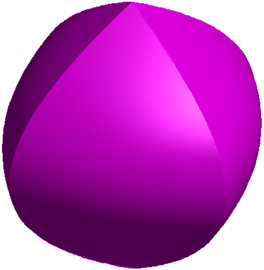
Copyright G. Obozinski



What about more complicated norms?

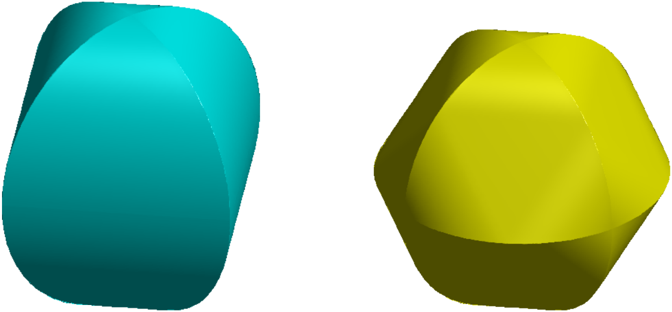
Copyright G. Obozinski





What about more complicated norms?

Copyright G. Obozinski



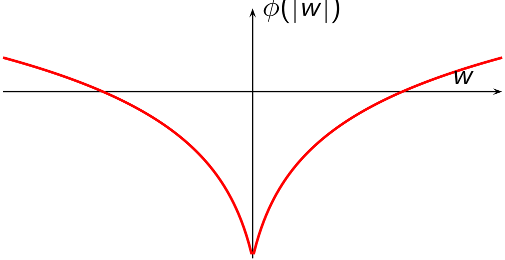
Examples of sparsity-inducing penalties

Exploiting concave functions with a kink at zero

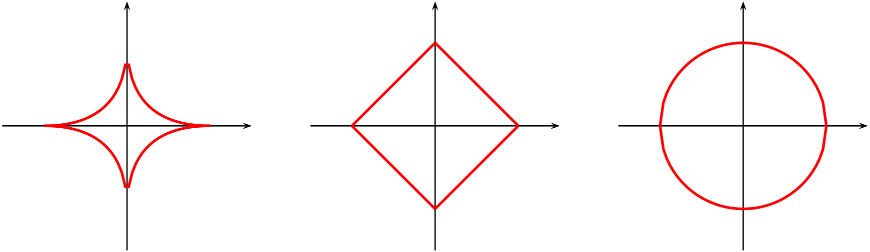
$\psi(\mathbf{w}) = \sum_{i=1}^{p} \phi(|\mathbf{w}_i|)$.

- ℓ_q-"pseudo-norm", with $0 < q < 1$: $\psi(\mathbf{w}) \triangleq \sum_{i=1}^{p} |\mathbf{w}_i|^q$,
- log penalty, $\psi(\mathbf{w}) \triangleq \sum_{i=1}^{p} \log(|\mathbf{w}_i| + \varepsilon)$,

ϕ is any function that looks like this:



Examples of sparsity-inducing penalties



(c) $\ell_{0.5}$-ball, 2-D (d) ℓ_1-ball, 2-D (e) ℓ_2-ball, 2-D

Figure: Open balls in 2-D corresponding to several ℓ_q-norms and pseudo-norms.

Examples of sparsity-inducing penalties

- The ℓ_1-ℓ_2 norm (group Lasso),

$$\sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2 = \sum_{g \in \mathcal{G}} \Big(\sum_{j \in g} \mathbf{w}_j^2 \Big)^{1/2}, \text{ with } \mathcal{G} \text{ a } \textbf{partition} \text{ of } \{1, \dots, p\}.$$

  selects groups of variables [Yuan and Lin, 2006].

- the fused Lasso [Tibshirani et al., 2005] or total variation [Rudin et al., 1992]:

$$\psi(\mathbf{w}) = \sum_{j=1}^{p-1} |\mathbf{w}_{j+1} - \mathbf{w}_j|.$$

Extensions (out of the scope of this talk):

- hierarchical norms [Zhao et al., 2009].
- structured sparsity [Jenatton et al., 2009, Jacob et al., 2009, Huang et al., 2009, Baraniuk et al., 2010]

Part II: Dictionary Learning and Matrix Factorization

Matrix Factorization and Clustering

Let us cluster some training vectors $\mathbf{y}^1, \ldots, \mathbf{y}^m$ into p clusters using K-means:

$$\min_{(\mathbf{x}^j)_{j=1}^p, (l_i)_{i=1}^m} \sum_{i=1}^m \|\mathbf{y}^i - \mathbf{x}^{l_i}\|_2^2.$$

It can be equivalently formulated as

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times p}, \mathbf{W} \in \{0,1\}^{p \times m}} \sum_{i=1}^m \|\mathbf{y}^i - \mathbf{X}\mathbf{w}^i\|_F^2 \;\; \text{s.t.} \;\; \mathbf{w}^i \geq 0 \text{ and } \sum_{j=1}^p \mathbf{w}_j^i = 1,$$

which is a **matrix factorization** problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times p}, \mathbf{W} \in \{0,1\}^{p \times m}} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 \;\; \text{s.t.} \;\; \mathbf{W} \geq 0 \text{ and } \sum_{j=1}^p \mathbf{w}_j^i = 1,$$

Matrix Factorization and Clustering

Hard clustering

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p},\mathbf{W}\in\{0,1\}^{p\times m}} \|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2 \ \text{ s.t. } \ \mathbf{W}\geq 0 \text{ and } \sum_{j=1}^{p}\mathbf{w}_j^i = 1,$$

$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ are the centroids of the p clusters.

Soft clustering

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p},\mathbf{W}\in\mathbb{R}^{p\times m}} \|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2 \ \text{ s.t. } \ \mathbf{W}\geq 0 \text{ and } \sum_{j=1}^{p}\mathbf{w}_j^i = 1,$$

$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ are the centroids of the p clusters.

Other Matrix Factorization Problems
PCA

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times n}\\ \mathbf{X}\in\mathbb{R}^{m\times p}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2 \;\text{ s.t. }\; \mathbf{X}^\top\mathbf{X}=\mathbf{I} \text{ and } \mathbf{W}\mathbf{W}^\top \text{ is diagonal.}$$

$\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_p]$ are the principal components.

Other Matrix Factorization Problems

Non-negative matrix factorization [Lee and Seung, 2001]

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times n}\\ \mathbf{X}\in\mathbb{R}^{m\times p}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2 \;\; \text{s.t.} \;\; \mathbf{W}\geq 0 \text{ and } \mathbf{X}\geq 0.$$

Dictionary Learning and Matrix Factorization

[Olshausen and Field, 1997]

$$\min_{\mathbf{X}\in\mathcal{X},\mathbf{W}\in\mathbb{R}^{p\times m}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{y}^i - \mathbf{X}\mathbf{w}^i\|_F^2 + \lambda\|\mathbf{w}^i\|_1,$$

which is again a matrix factorization problem

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times n}\\ \mathbf{X}\in\mathcal{X}}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1.$$

Why having a unified point of view?

$$\min_{\substack{\mathbf{W}\in\mathcal{W}\\ \mathbf{X}\in\mathcal{X}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2 + \lambda\psi(\mathbf{W}).$$

- same framework for NMF, sparse PCA, dictionary learning, clustering, topic modelling;
- can play with various constraints/penalties on $\mathbf{W}$ (coefficients) and on $\mathbf{X}$ (loadings, dictionary, centroids);
- same algorithms (no need to reinvent the wheel): alternate minimization, online learning [Mairal et al., 2010].

Advertisement SPAMS toolbox (open-source)

- C++ interfaced with **Matlab, R, Python**.
- proximal gradient methods for ℓ_0, ℓ_1, **elastic-net, fused-Lasso, group-Lasso, tree group-Lasso, tree-ℓ_0, sparse group Lasso, overlapping group Lasso...**
- ...for **square, logistic, multi-class logistic** loss functions.
- handles sparse matrices, provides duality gaps.
- fast implementations of **OMP** and **LARS - homotopy**.
- dictionary learning and matrix factorization (NMF, sparse PCA).
- coordinate descent, block coordinate descent algorithms.
- fast projections onto some convex sets.

Try it! `http://www.di.ens.fr/willow/SPAMS/`

Part III: A few Image Processing Stories

The Image Denoising Problem

$$\underbrace{\mathbf{y}}_{\text{measurements}} = \underbrace{\mathbf{x}_{orig}}_{\text{original image}} + \underbrace{\mathbf{w}}_{\text{noise}}$$

Sparse representations for image restoration

$$\underbrace{\mathbf{y}}_{\text{measurements}} = \underbrace{\mathbf{x}_{orig}}_{\text{original image}} + \underbrace{\mathbf{w}}_{\text{noise}}$$

Energy minimization problem - MAP estimation

$$E(\mathbf{x}) = \underbrace{\frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2}_{\text{relation to measurements}} + \underbrace{\psi(\mathbf{x})}_{\text{image model (-log prior)}}$$

Some classical priors

- Smoothness $\lambda\|\mathcal{L}\mathbf{x}\|_2^2$
- Total variation $\lambda\|\nabla\mathbf{x}\|_1^2$
- MRF priors
- ...

Sparse representations for image restoration

Designed dictionaries

[Haar, 1910], [Zweig, Morlet, Grossman ∼70s], [Meyer, Mallat, Daubechies, Coifman, Donoho, Candes ∼80s-today]. . . (see [Mallat, 1999])
Wavelets, Curvelets, Wedgelets, Bandlets, . . . lets

Learned dictionaries of patches

[Olshausen and Field, 1997], [Engan et al., 1999], [Lewicki and Sejnowski, 2000], [Aharon et al., 2006] , [Roth and Black, 2005], [Lee et al., 2007]

$$\min_{\mathbf{w}_i, \mathbf{X} \in \mathcal{C}} \sum_i \underbrace{\frac{1}{2}\|\mathbf{y}_i - \mathbf{X}\mathbf{w}_i\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda\psi(\mathbf{w}_i)}_{\text{sparsity}}$$

- $\psi(\mathbf{w}) = \|\mathbf{w}\|_0$ (“ℓ_0 pseudo-norm”)
- $\psi(\mathbf{w}) = \|\mathbf{w}\|_1$ (ℓ_1 norm)

Sparse representations for image restoration

Solving the denoising problem

[Elad and Aharon, 2006]

- Extract all overlapping 8×8 patches $\mathbf{y}_i$.
- Solve a matrix factorization problem:

$$\min_{\mathbf{w}_i, \mathbf{X} \in \mathcal{C}} \sum_{i=1}^{n} \underbrace{\frac{1}{2}\|\mathbf{y}_i - \mathbf{X}\mathbf{w}_i\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda\psi(\mathbf{w}_i)}_{\text{sparsity}},$$

  with $n > 100,000$

- Average the reconstruction of each patch.

Sparse representations for image restoration

K-SVD: [Elad and Aharon, 2006]

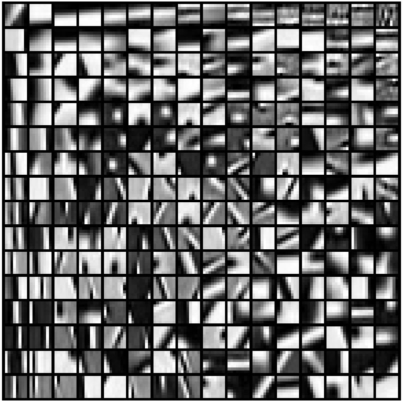
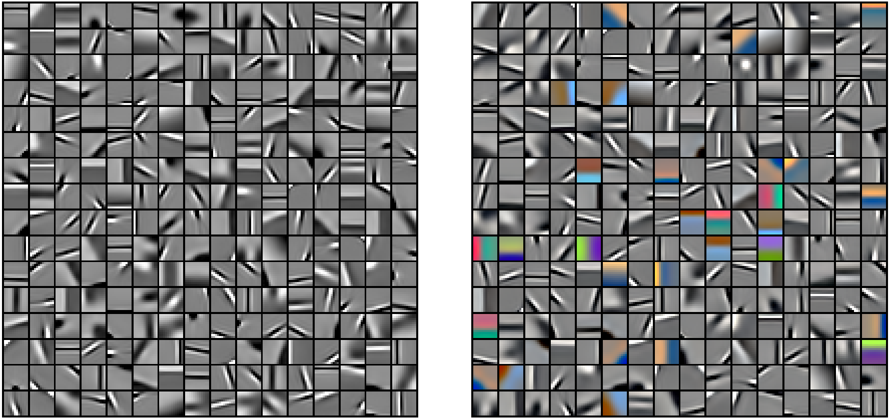


Figure: Dictionary trained on a noisy version of the image boat.

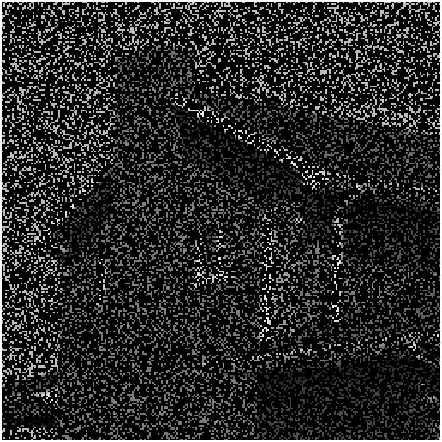
Sparse representations for image restoration

Grayscale vs color image patches



Sparse representations for image restoration

[Mairal, Sapiro, and Elad, 2008b]



Sparse representations for image restoration

Inpainting, [Mairal, Elad, and Sapiro, 2008a]

Sparse representations for image restoration

Inpainting, [Mairal, Elad, and Sapiro, 2008a]

Sparse representations for video restoration

Key ideas for video processing

[Protter and Elad, 2009]

- Using a 3D dictionary.
- Processing of many frames at the same time.
- Dictionary propagation.

Sparse representations for image restoration
Inpainting, [Mairal, Sapiro, and Elad, 2008b]

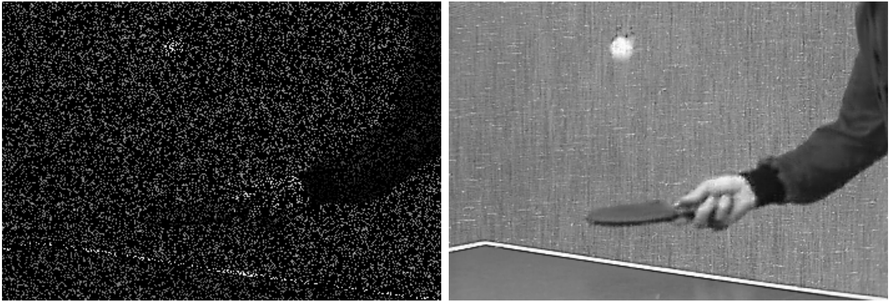


Figure: Inpainting results.

Sparse representations for image restoration
Inpainting, [Mairal, Sapiro, and Elad, 2008b]

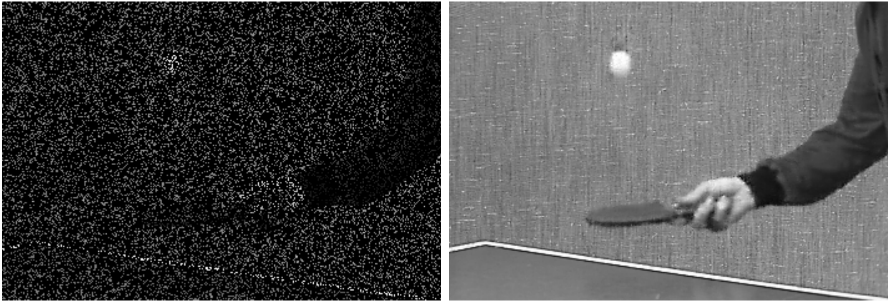


Figure: Inpainting results.

Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]

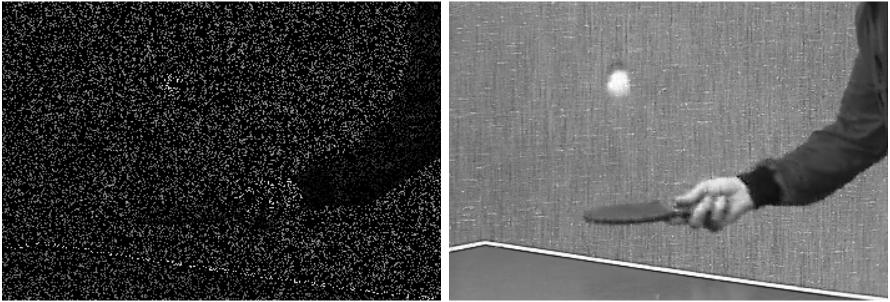


Figure: Inpainting results.

Sparse representations for image restoration
Inpainting, [Mairal, Sapiro, and Elad, 2008b]

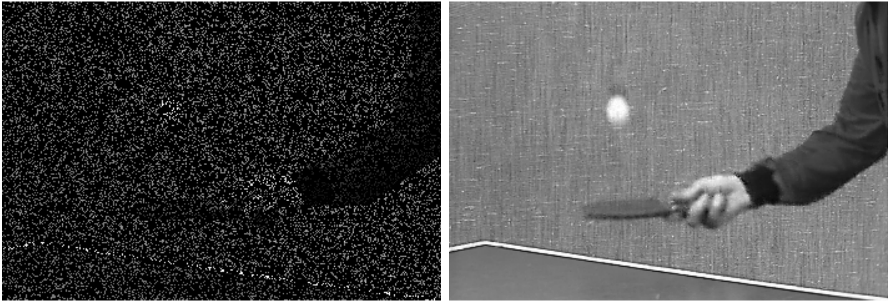


Figure: Inpainting results.

Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]

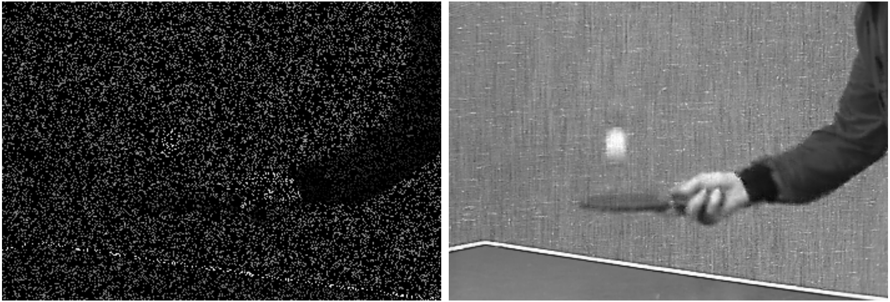


Figure: Inpainting results.

Sparse representations for image restoration
Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

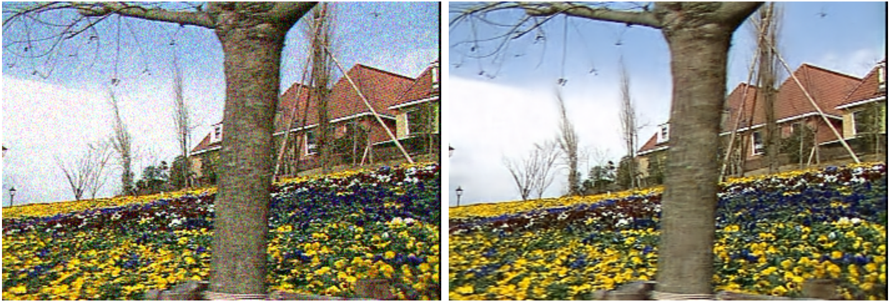


Figure: Denoising results. $\sigma = 25$

Sparse representations for image restoration
Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Sparse representations for image restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Sparse representations for image restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Sparse representations for image restoration
Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Digital Zooming

Couzinie-Devy, 2010, Original

Digital Zooming

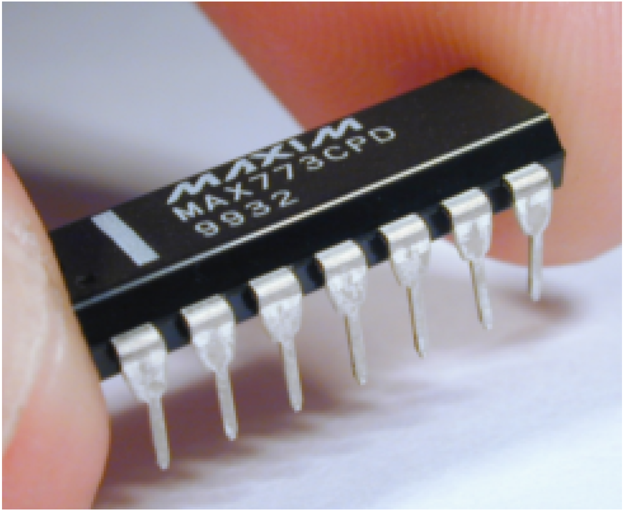
Couzinie-Devy, 2010, Bicubic

Digital Zooming

Couzinie-Devy, 2010, Proposed method

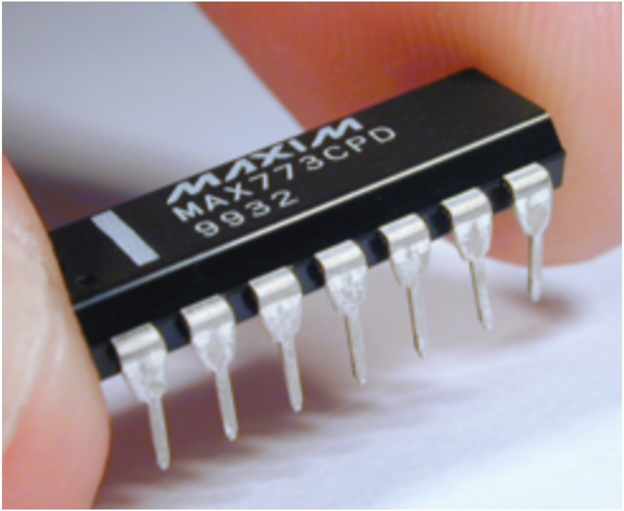
Digital Zooming

Couzinie-Devy, 2010, Original



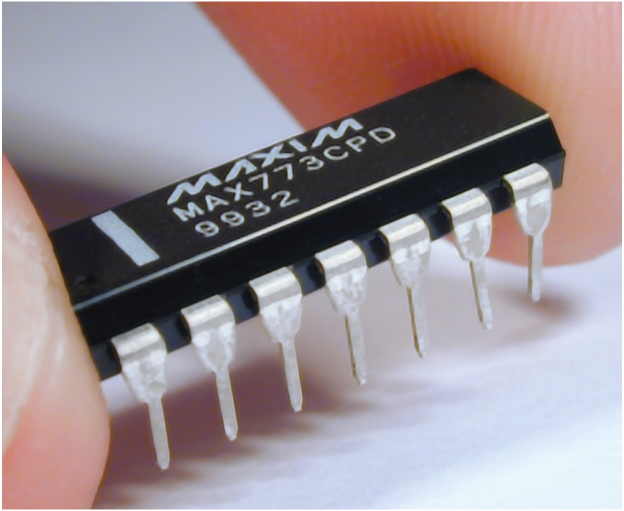
Digital Zooming

Couzinie-Devy, 2010, Bicubic



Digital Zooming

Couzinie-Devy, 2010, Proposed approach



Inverse half-toning

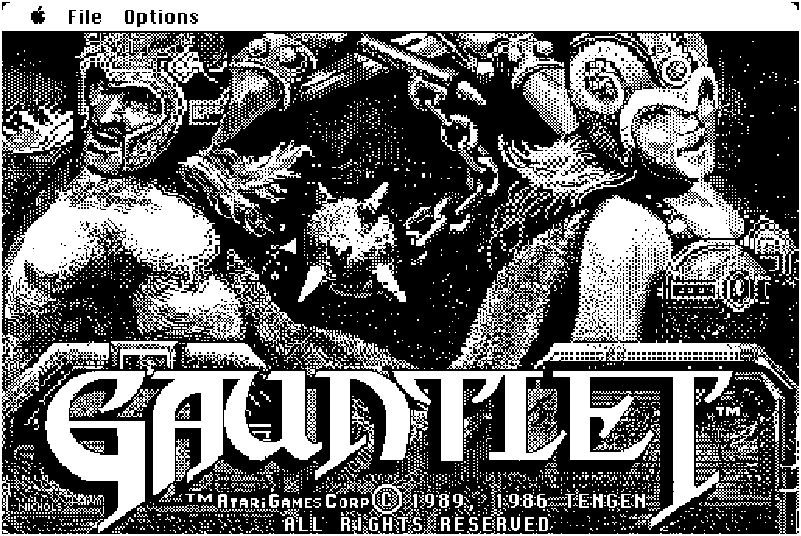
Original

Inverse half-toning

Reconstructed image

Inverse half-toning

Original



Inverse half-toning

Reconstructed image

Inverse half-toning

Original

Inverse half-toning

Reconstructed image

Inverse half-toning

Original

Inverse half-toning

Reconstructed image

Inverse half-toning

Original

Inverse half-toning

Reconstructed image

Conclusion

- We have seen that many formulations are related to sparse regularized matrix factorization problems: pca, sparse pca, clustering, nmf, dictionary learning;
- we have so successful stories in images processing, computer vision and neuroscience;
- there exists efficient software for Matlab/R/Python.

References I

M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, November 2006.

R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 2010. to appear.

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.

E. Candes. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, 2006.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1999.

References II

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 54(12):3736–3745, December 2006.

K. Engan, S. O. Aase, and J. H. Husoy. Frame based signal compression using method of optimal directions (MOD). In *Proceedings of the 1999 IEEE International Symposium on Circuits Systems*, volume 4, 1999.

A. Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69:331–371, 1910.

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

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

References III

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, 2009. preprint arXiv:0904.3523v1.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 2001.

H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 801–808. MIT Press, Cambridge, MA, 2007.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, January 2008a.

References IV

J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *SIAM Multiscale Modelling and Simulation*, 7(1):214–241, April 2008b.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 2010.

S. Mallat. *A Wavelet Tour of Signal Processing, Second Edition*. Academic Press, New York, September 1999.

S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

Y. Nesterov. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 27:372–376, 1983.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, CORE, 2007.

References V

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37: 3311–3325, 1997.

M. Protter and M. Elad. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing*, 18(1):27–36, 2009.

S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4): 259–268, 1992.

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

References VI

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, 67(1):91–108, 2005.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 2006.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. 37(6A):3468–3497, 2009.

One short slide on compressed sensing

Important message

Sparse coding is not "compressed sensing".

Compressed sensing is a theory [see Candes, 2006] saying that a sparse signal can be recovered from a few linear measurements under some conditions.

- Signal Acquisition: $\mathbf{Z}^\top \mathbf{y}$, where $\mathbf{Z} \in \mathbb{R}^{m \times s}$ is a "sensing" matrix with $s \ll m$.
- Signal Decoding: $\min_{\mathbf{w} \in \mathbb{R}^p} \|\mathbf{w}\|_1$ s.t. $\mathbf{Z}^\top \mathbf{y} = \mathbf{Z}^\top \mathbf{X} \mathbf{w}$.

with extensions to approximately sparse signals, noisy measurements.

Remark

The dictionaries we are using in this lecture do not satisfy the recovery assumptions of compressed sensing.

Greedy Algorithms

Several equivalent non-convex and NP-hard problems:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2}_{\text{residual } \mathbf{r}} + \underbrace{\lambda\|\mathbf{w}\|_0}_{\text{regularization}} ,$$

$$\min_{\mathbf{w}\in\mathbb{R}^p} \|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 \ \text{ s.t. } \ \|\mathbf{w}\|_0 \leq L,$$

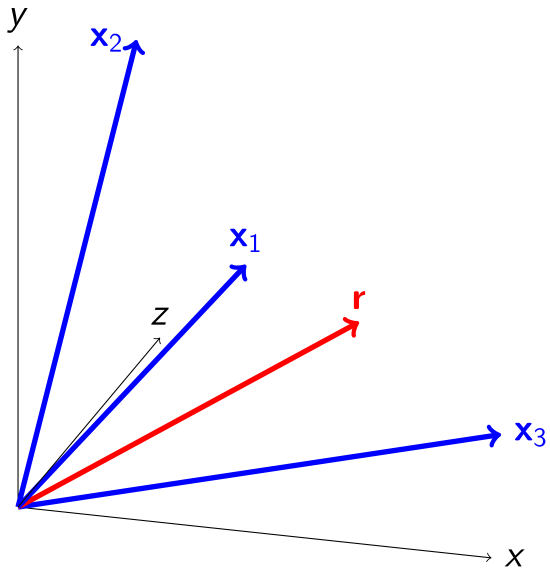
$$\min_{\mathbf{w}\in\mathbb{R}^p} \|\mathbf{w}\|_0 \ \text{ s.t. } \ \|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 \leq \varepsilon,$$

The solution is often approximated with a **greedy** algorithm.

- **Signal processing**: Matching Pursuit [Mallat and Zhang, 1993], Orthogonal Matching Pursuit [**?**].
- **Statistics**: L2-boosting, forward selection.

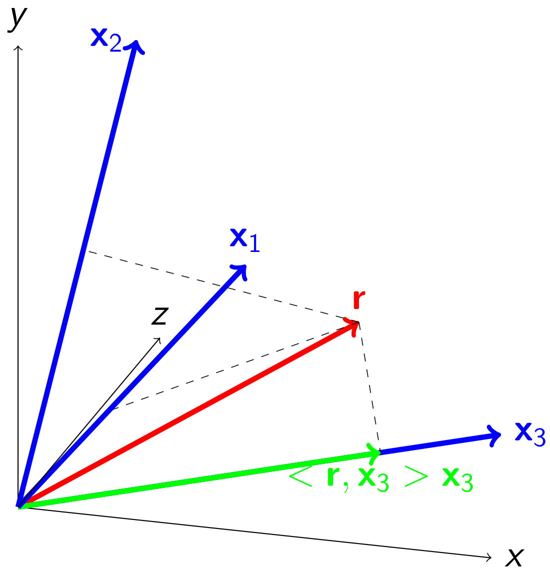
Matching Pursuit

$$\mathbf{w} = (0, 0, 0)$$



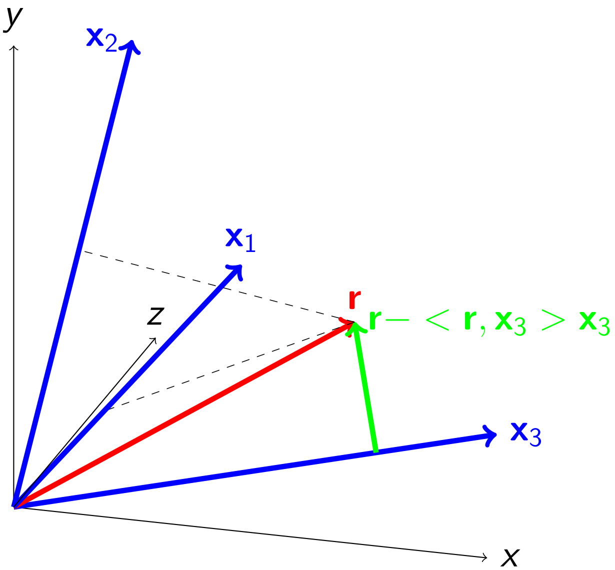
Matching Pursuit

$$\mathbf{w} = (0, 0, 0)$$



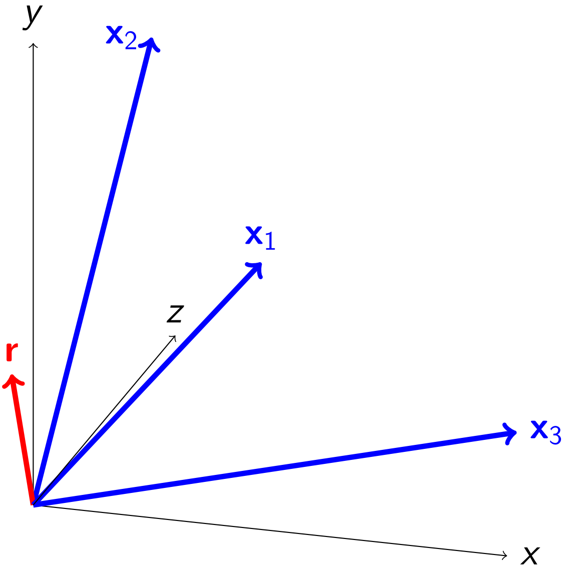
Matching Pursuit

$$\mathbf{w} = (0, 0, 0)$$



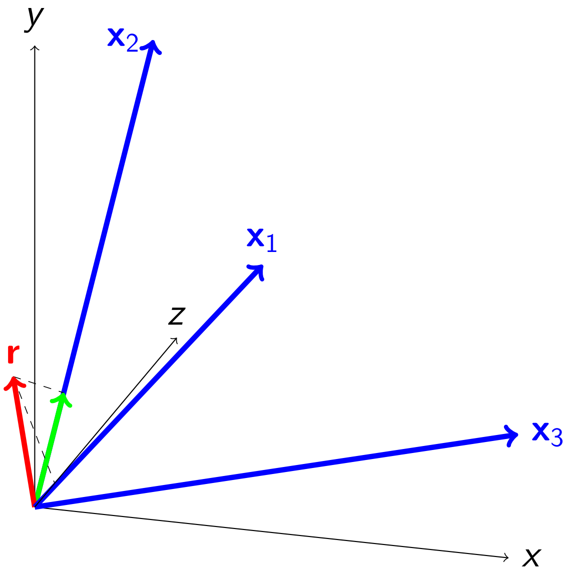
Matching Pursuit

$$\mathbf{w} = (0, 0, 0.75)$$



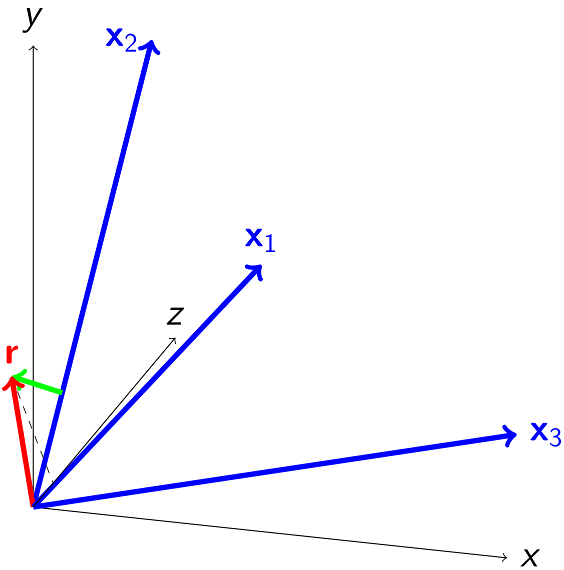
Matching Pursuit

$$\mathbf{w} = (0, 0, 0.75)$$



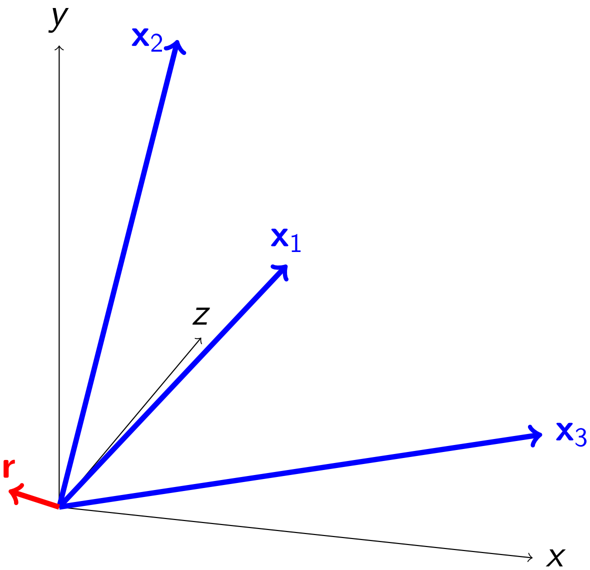
Matching Pursuit

$$\mathbf{w} = (0, 0, 0.75)$$



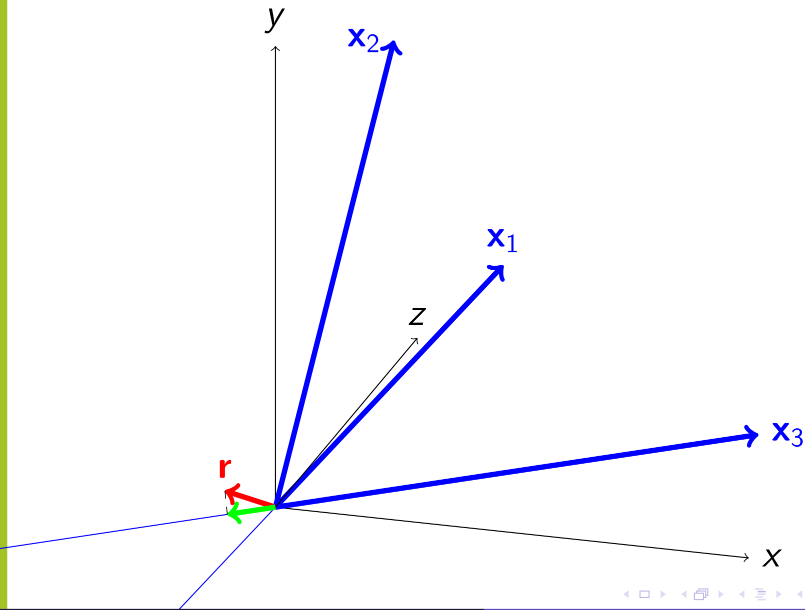
Matching Pursuit

$$\mathbf{w} = (0, 0.24, 0.75)$$



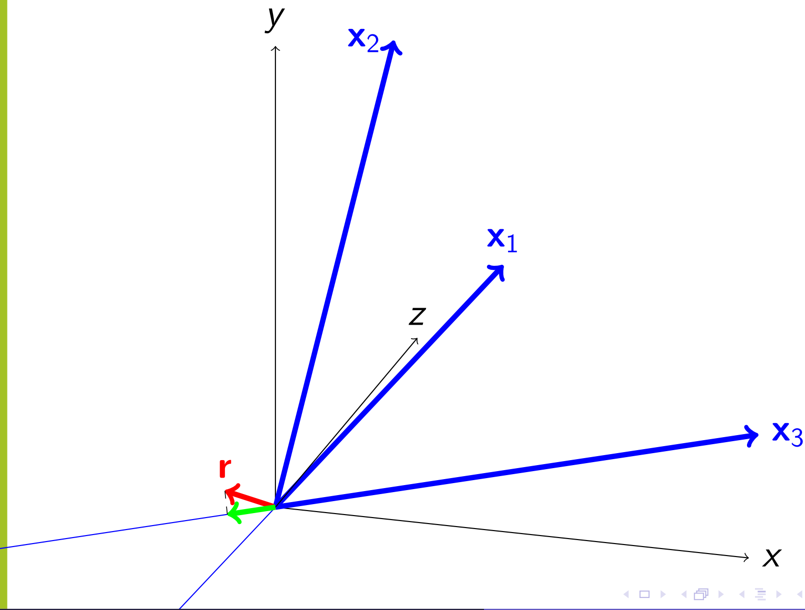
Matching Pursuit

$$\mathbf{w} = (0, 0.24, 0.75)$$



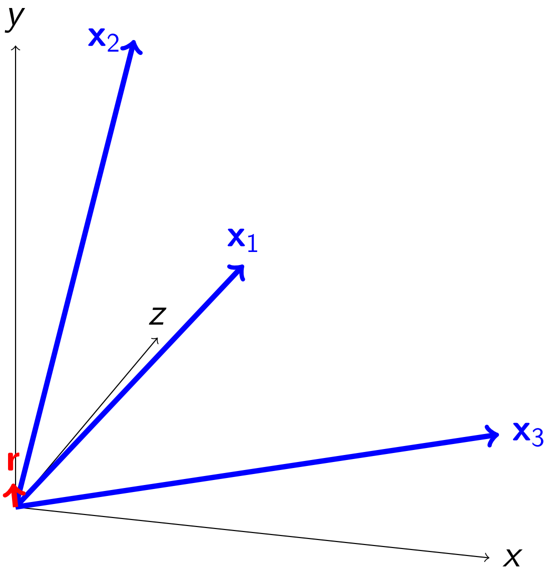
Matching Pursuit

$$\mathbf{w} = (0, 0.24, 0.75)$$



Matching Pursuit

$$\mathbf{w} = (0, 0.24, 0.65)$$



Matching Pursuit

$$\min_{\mathbf{w}\in\mathbb{R}^p} \|\underbrace{\mathbf{y}-\mathbf{X}\mathbf{w}}_{\mathbf{r}}\|_2^2 \;\; \text{s.t.} \;\; \|\mathbf{w}\|_0 \leq L$$

$\mathbf{w} \leftarrow 0$

$\mathbf{r} \leftarrow \mathbf{y}$ (residual).

while $\|\mathbf{w}\|_0 < L$ **do**

  Select the predictor with maximum correlation with the residual

$$\hat{\imath} \leftarrow \arg\max_{i=1,\ldots,p} |\mathbf{x}^{i\top}\mathbf{r}|$$

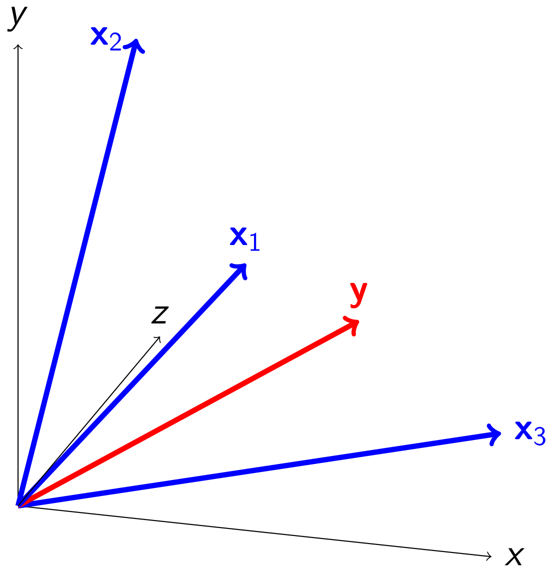
  Update the residual and the coefficients

$$\begin{aligned} \mathbf{w}_{\hat{\imath}} &\leftarrow \mathbf{w}_{\hat{\imath}} + \mathbf{x}^{\hat{\imath}\top}\mathbf{r} \\ \mathbf{r} &\leftarrow \mathbf{r} - (\mathbf{x}^{\hat{\imath}\top}\mathbf{r})\mathbf{x}^{\hat{\imath}} \end{aligned}$$

end while

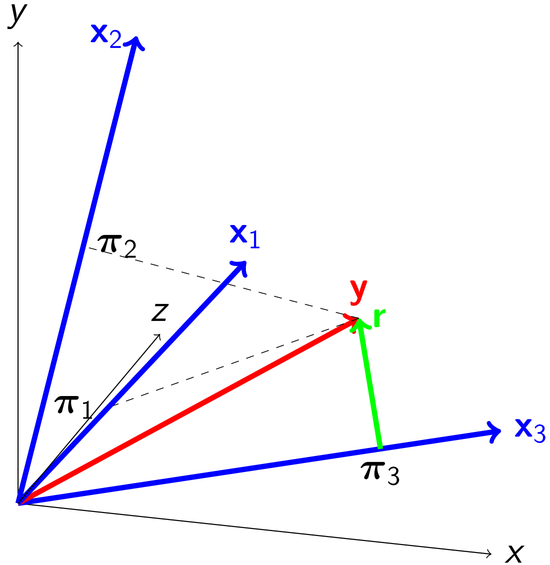
Orthogonal Matching Pursuit

$$\mathbf{w} = (0,0,0)$$
$$\mathsf{J} = \emptyset$$



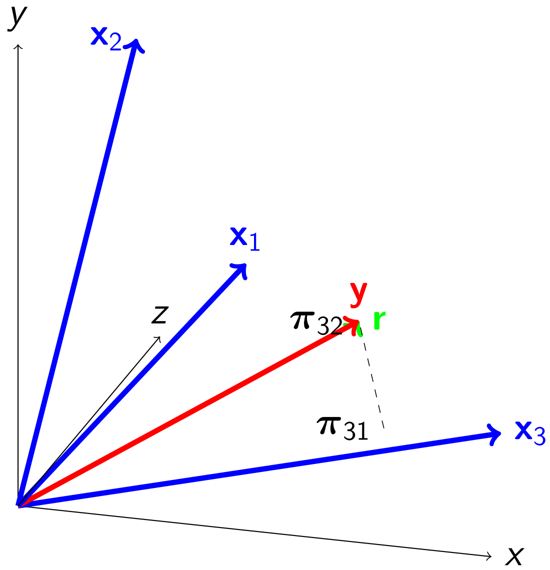
Orthogonal Matching Pursuit

$$\mathbf{w} = (0, 0, 0.75)$$
$$\mathsf{J} = \{3\}$$



Orthogonal Matching Pursuit

$$\mathbf{w} = (0, 0.29, 0.63)$$
$$J = \{3, 2\}$$



Orthogonal Matching Pursuit

$$\min_{\mathbf{w}\in\mathbb{R}^p} \|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 \;\text{ s.t. }\; \|\mathbf{w}\|_0 \leq L$$

1: $\mathsf{J} = \emptyset$.
2: **for** *iter* $= 1, \ldots, L$ **do**
3: Select the predictor which most reduces the objective

$$\hat{\imath} \leftarrow \arg\min_{i\in\mathsf{J}^C} \left\{ \min_{\mathbf{w}'} \|\mathbf{y}-\mathbf{X}_{\mathsf{J}\cup\{i\}}\mathbf{w}'\|_2^2 \right\}$$

4: Update the active set: $\mathsf{J} \leftarrow \mathsf{J} \cup \{\hat{\imath}\}$.
5: Update the residual (orthogonal projection)

$$\mathbf{r} \leftarrow (\mathbf{I} - \mathbf{X}_\mathsf{J}(\mathbf{X}_\mathsf{J}^\top\mathbf{X}_\mathsf{J})^{-1}\mathbf{X}_\mathsf{J}^\top)\mathbf{y}.$$

6: Update the coefficients

$$\mathbf{w}_\mathsf{J} \leftarrow (\mathbf{X}_\mathsf{J}^\top\mathbf{X}_\mathsf{J})^{-1}\mathbf{X}_\mathsf{J}^\top\mathbf{y}.$$

7: **end for**

Orthogonal Matching Pursuit

The keys for a good implementation

- If available, use Gram matrix $\mathbf{G} = \mathbf{X}^\top \mathbf{X}$,
- Maintain the computation of $\mathbf{X}^\top \mathbf{r}$ for each signal,
- Maintain a Cholesky decomposition of $(\mathbf{X}_\mathsf{J}^\top \mathbf{X}_\mathsf{J})^{-1}$ for each signal.

The total complexity for decomposing n L-sparse signals of size m with a dictionary of size p is

$$\underbrace{O(p^2 m)}_{\text{Gram matrix}} + \underbrace{O(nL^3)}_{\text{Cholesky}} + \underbrace{O(n(pm + pL^2))}_{\mathbf{X}^T \mathbf{r}} = O(np(m + L^2))$$

It is also possible to use the matrix inversion lemma instead of a Cholesky decomposition (same complexity, but less numerical stability).

Coordinate Descent for the Lasso

- Coordinate descent + nonsmooth objective: **WARNING: not convergent in general**
- Here, the problem is equivalent to a convex smooth optimization problem with **separable** constraints

$$\min_{\mathbf{w}_+,\mathbf{w}_-} \frac{1}{2}\|\mathbf{y} - \mathbf{X}_+\mathbf{w}_+ + \mathbf{X}_-\mathbf{w}_-\|_2^2 + \lambda\mathbf{w}_+^T\mathbf{1} + \lambda\mathbf{w}_-^T\mathbf{1} \;\text{ s.t. }\; \mathbf{w}_-,\mathbf{w}_+ \geq 0.$$

- For this **specific** problem, coordinate descent is **convergent**.
- Assume the colums of $\mathbf{X}$ to have unit ℓ_2-norm, updating the coordinate i:

$$\mathbf{w}_i \leftarrow \arg\min_{w \in \mathbb{R}} \frac{1}{2}\|\underbrace{\mathbf{y} - \sum_{j \neq i} \mathbf{w}_j \mathbf{x}^j}_{\mathbf{r}} - w\mathbf{x}^i\|_2^2 + \lambda|w|$$

$$\leftarrow \operatorname{sign}(\mathbf{x}^{i\top}\mathbf{r})(|\mathbf{x}^{i\top}\mathbf{r}| - \lambda)^+$$

- $\Rightarrow$ **soft-thresholding**!

First-order/proximal methods

$$\min_{\mathbf{w}\in\mathbb{R}^p} f(\mathbf{w}) + \lambda\psi(\mathbf{w})$$

- f is strictly convex and differentiable with a Lipshitz gradient.
- Generalizes the idea of gradient descent

$$\mathbf{w}^{k+1} \leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{f(\mathbf{w}^k) + \nabla f(\mathbf{w}^k)^\top(\mathbf{w}-\mathbf{w}^k)}_{\text{linear approximation}} + \underbrace{\frac{L}{2}\|\mathbf{w}-\mathbf{w}^k\|_2^2}_{\text{quadratic term}} + \lambda\psi(\mathbf{w})$$

$$\leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{w} - (\mathbf{w}^k - \frac{1}{L}\nabla f(\mathbf{w}^k))\|_2^2 + \frac{\lambda}{L}\psi(\mathbf{w})$$

When $\lambda = 0$, $\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \frac{1}{L}\nabla f(\mathbf{w}^k)$, this is equivalent to a classical gradient descent step.

First-order/proximal methods

- They require solving efficiently the proximal operator

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2 + \lambda\psi(\mathbf{w})$$

- For the ℓ_1-norm, this amounts to a soft-thresholding:

$$\mathbf{w}_i^\star = \mathrm{sign}(\mathbf{u}_i)(\mathbf{u}_i - \lambda)^+.$$

- There exists accelerated versions based on Nesterov optimal first-order method (gradient method with "extrapolation") [Beck and Teboulle, 2009, Nesterov, 2007, 1983]
- suited for large-scale experiments.

Optimization for Grouped Sparsity

The proximal operator:

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\|_2^2 + \lambda \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_q$$

For $q = 2$,

$$\mathbf{w}_g^\star = \frac{\mathbf{u}_g}{\|\mathbf{u}_g\|_2}(\|\mathbf{u}_g\|_2 - \lambda)^+, \quad \forall g \in \mathcal{G}$$

For $q = \infty$,

$$\mathbf{w}_g^\star = \mathbf{u}_g - \Pi_{\|.\|_1 \leq \lambda}[\mathbf{u}_g], \quad \forall g \in \mathcal{G}$$

These formula generalize soft-thrsholding to groups of variables. They are used in block-coordinate descent and proximal algorithms.

Smoothing Techniques: Reweighted ℓ_2

Let us start from something simple

$$a^2 - 2ab + b^2 \geq 0.$$

Then

$$a \leq \frac{1}{2}\left(\frac{a^2}{b} + b\right) \text{ with equality iff } a = b$$

and

$$\|\mathbf{w}\|_1 = \min_{\eta_j \geq 0} \frac{1}{2} \sum_{j=1}^{p} \frac{\mathbf{w}[j]^2}{\eta_j} + \eta_j.$$

The formulation becomes

$$\min_{\mathbf{w}, \eta_j \geq \varepsilon} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{2} \sum_{j=1}^{p} \frac{\mathbf{w}[j]^2}{\eta_j} + \eta_j.$$

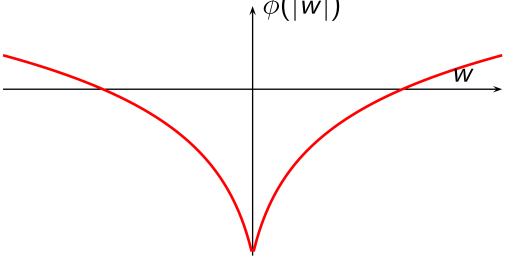
DC (difference of convex) - Programming

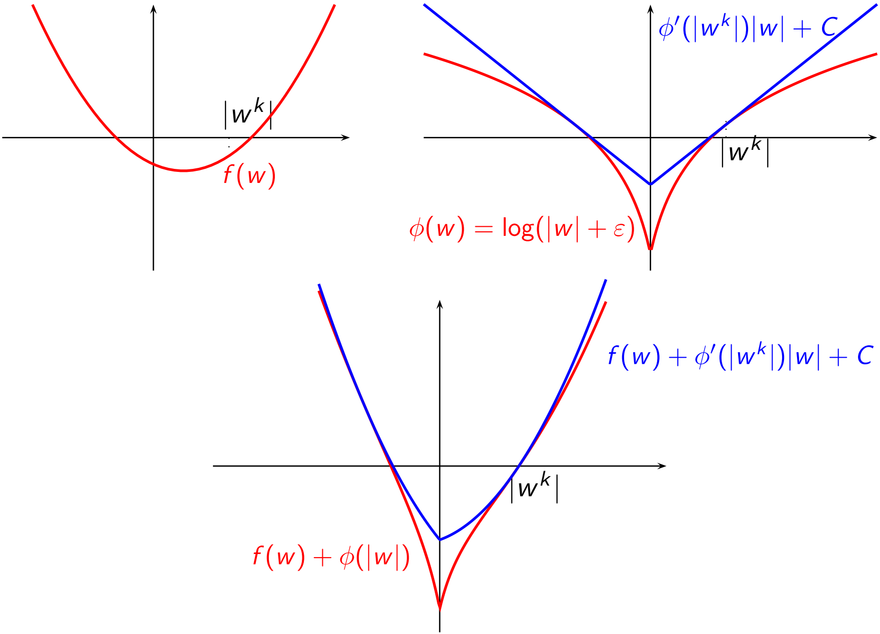
Remember? Concave functions with a kink at zero

$\psi(\mathbf{w}) = \sum_{i=1}^{p} \phi(|\mathbf{w}_i|)$.

- ℓ_q-"pseudo-norm", with $0 < q < 1$: $\psi(\mathbf{w}) \triangleq \sum_{i=1}^{p} |\mathbf{w}_i|^q$,
- log penalty, $\psi(\mathbf{w}) \triangleq \sum_{i=1}^{p} \log(|\mathbf{w}_i| + \varepsilon)$,

ϕ is any function that looks like this:





$|w^k|$
$f(w)$
$\phi'(|w^k|)|w| + C$
$|w^k|$
$\phi(w) = \log(|w| + \varepsilon)$
$f(w) + \phi'(|w^k|)|w| + C$
$|w^k|$
$f(w) + \phi(|w|)$

DC (difference of convex) - Programming

$$\min_{\mathbf{w} \in \mathbb{R}^p} f(\mathbf{w}) + \lambda \sum_{i=1}^{p} \phi(|\mathbf{w}_i|).$$

This problem is non-convex. f is convex, and ϕ is concave on $\mathbb{R}^+$.
if $\mathbf{w}^k$ is the current estimate at iteration k, the algorithm solves

$$\mathbf{w}^{k+1} \leftarrow \arg\min_{\mathbf{w} \in \mathbb{R}^p} \left[f(\mathbf{w}) + \lambda \sum_{i=1}^{p} \psi'(|\mathbf{w}_i^k|)|\mathbf{w}_i| \right],$$

which is a **reweighted**-ℓ_1 problem.
Warning: It does not solve the non-convex problem, only provides a stationary point.
In practice, each iteration sets to zero small coefficients. After $2-3$ iterations, the result does not change much.