Structure and Regularization

Neal Parikh

Computer Science Department

Stanford University

February 2013

Outline

Structure in variables

- often know or assume that solution to a problem is structured, *e.g.*,
  - convex-cardinality problems
  - high-dimensional statistics: assume low-dimensional structure
  - prior knowledge that variables have, *e.g.*, hierarchical or grouped structure

- handle by solving a problem with two conceptual components:
  - main objective of interest (model fit, satisfying constraints, . . .)
  - regularization term that encourages assumed form of structure

- possible structure of interest includes sparsity, low rank, . . .

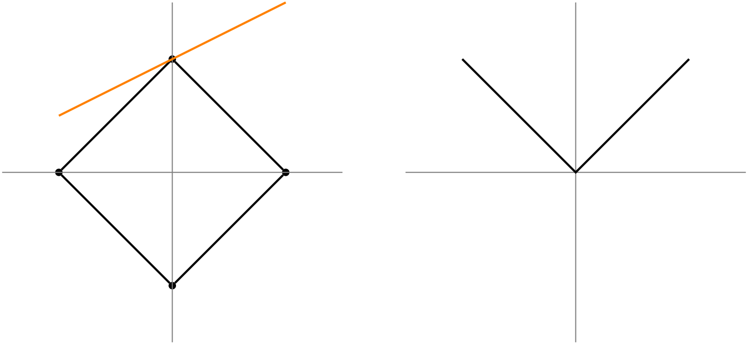
this talk:

❶ selecting regularization to promote assumed structure

❷ many examples and applications (*i.e.*, sparsify everything in sight)

❸ solving the resulting optimization problems

Geometric interpretation

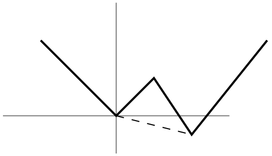


get sparsity/structure when corners/kinks appear at sparse/structured points

e.g., quadratic cone, linear functions on prob. simplex, nuclear norm, ...

Convex envelope interpretation

- convex envelope of (nonconvex) f is the largest convex underestimator g
- *i.e.*, the best convex lower bound to a function



- **example**: ℓ_1 is the envelope of $\mathbf{card}$ (on unit ℓ_∞ ball)
- **example**: $\|\cdot\|_*$ is the envelope of $\mathbf{rank}$ (on unit spectral norm ball)
- various characterizations: *e.g.*, f^{**} or convex hull of epigraph

Penalty function interpretation

- compared to ridge penalty $\|x\|_2^2$, using ℓ_1 does two things:

  ➊ higher emphasis on small values to go to exactly zero
  ➋ lower emphasis on avoiding very large values

- thus useful for obtaining **sparse** or **robust** solutions to problems

Atomic norm interpretation

(Chandrasekaran, Recht, Parrilo, Willsky)

- convex surrogates for measures of 'simplicity'
- suppose underlying parameter vector or signal $x \in \mathbf{R}^n$ given by

$$x = \sum_{i=1}^{k} c_i a_i, \quad a_i \in \mathcal{A},\ c_i \geq 0,$$

  where $\mathcal{A}$ is set of 'atoms' and $k \ll n$ (d.f. $\ll$ ambient dimension)

- if $\mathcal{A}$ is usual basis vectors, model says that x is k-sparse, and

$$\mathbf{conv}(\mathcal{A}) = \text{unit } \ell_1 \text{ ball}$$

- then, *e.g.*, minimize $\|x\|_1$ subject to $y = Fx$

Heuristics

- $\lambda_{\max}$ heuristic

  1. (analytically) compute $\lambda_{\max}$ as value for which $x^\star = 0$
  2. set $\lambda = \alpha\lambda_{\max}$, where $\alpha \in [0.01, 0.3]$

  e.g., for the lasso, $\lambda_{\max} = \|A^T b\|_\infty$

- polishing heuristic

  1. use ℓ_1 heuristic to find $\hat{x}$ with desired sparsity
  2. fix sparsity pattern
  3. re-solve (unregularized) problem with this pattern to obtain final solution

- reweighted ℓ_1 heuristic (Candès, Wakin, Boyd)

Outline

1. ℓ_1 regularization
2. Examples and extensions
3. Proximal algorithms
4. Conclusions

Sparse design

- find sparse design vector x satisfying specifications

$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

- zero values of x simplify design or correspond to unneeded components
- when $\mathcal{C} = \{x \mid Ax = b\}$, called **basis pursuit** or **sparse coding**
- *e.g.*, find sparse representation of signal b in 'dictionary' or 'overcomplete basis' given by columns of A

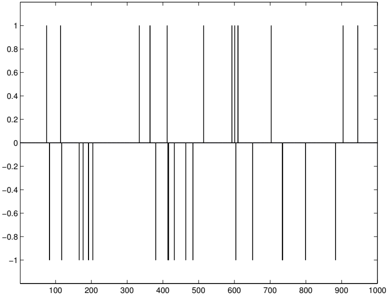
Sparse regression

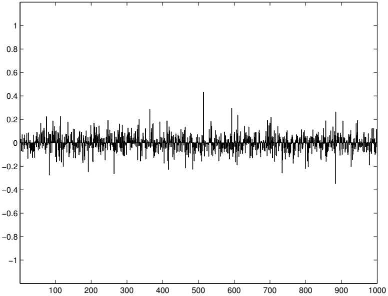
- fit $b \in \mathbf{R}^m$ as linear combination of a subset of regressors

$$\text{minimize} \quad (1/2)\|Ax - b\|_2^2 + \lambda\|x\|_1$$

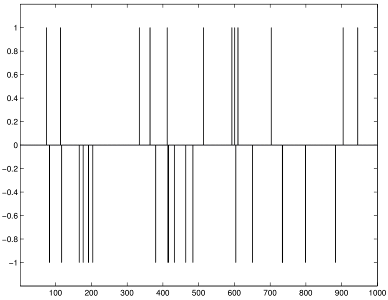
- zero values of x indicate features not predictive of the response
- also known as the lasso
- easily generalizes to other losses (*e.g.*, sparse logistic regression)

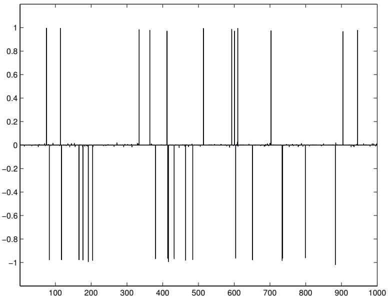
Sparse regression





Sparse regression





Estimation with outliers

- measurements $y_i = a_i^T x + v_i + w_i$
- v_i is Gaussian noise (small), w is a sparse outlier vector (big)
- if $\mathcal{O} = \{i \mid w_i \neq 0\}$ is set of outliers, MLE given by

$$\begin{array}{ll} \text{minimize} & \sum_{i \notin \mathcal{O}} (y_i - a_i^T x)^2 \\ \text{subject to} & |\mathcal{O}| \leq k \end{array}$$

- convex approximation given by

$$\text{minimize} \quad (1/2)\|y - Ax - w\|_2^2 + \lambda \|w\|_1$$

- same idea used in support vector machine

Linear classifier with fewest errors

- want linear classifier $b \approx \mathbf{sign}(a^T x + s)$ from $(a_i, b_i) \in \mathbf{R}^n \times \{-1, 1\}$
- error corresponds to negative margin: $b_i(a_i^T x + s) \leq 0$
- find x, s that give fewest classification errors:

$$\begin{array}{ll} \text{minimize} & \|t\|_1 \\ \text{subject to} & b_i(a_i^T x + s) + t_i \geq 1, \quad i = 1, \ldots m \end{array}$$

  with variables x, s, t

- close to a support vector machine
- can generalize to other convex feasibility problems

Elastic net

(Zou & Hastie)

- problem:

$$\text{minimize} \quad f(x) + \lambda\|x\|_1 + (1-\lambda)\|x\|_2^2$$

  i.e., use both ridge and lasso penalties

- attempts to overcome the following potential drawbacks of the lasso:
  - lasso selects at most (# examples) variables
  - given group of very correlated features, lasso often picks one arbitrarily

- here, strongly correlated predictors are jointly included or not

- (in practice, need to do some rescaling above)

Fused lasso

(Tibshirani et al.; Rudin, Osher, Fatemi)

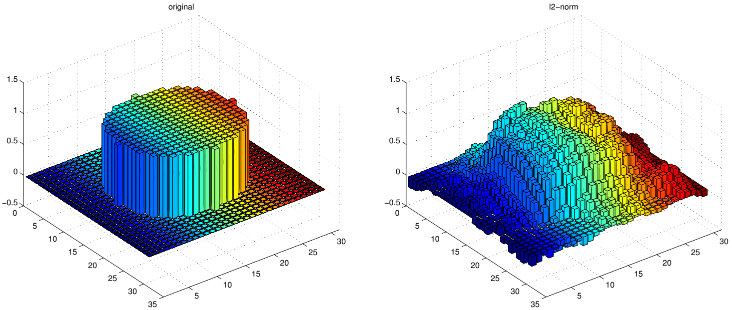
- problem:

$$\text{minimize} \quad f(x) + \lambda_1 \|x\|_1 + \lambda_2 \sum_{j=2}^{n} |x_j - x_{j-1}|$$

  i.e., encourage x to be both sparse and piecewise constant

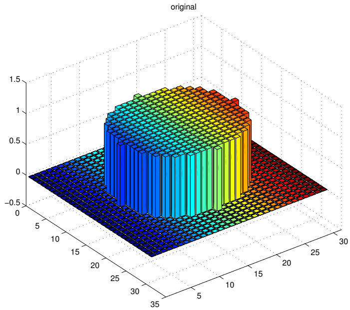
- special case: **total variation denoising** (set $\lambda_1 = 0$)
- used in biology (*e.g.*, gene expression) and signal reconstruction
- can also write penalty as $\|Dx\|_1$; could consider other matrices

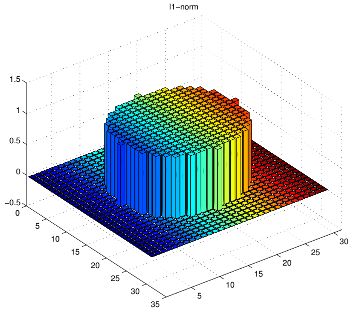
Total variation denoising



120 linear measurements and $31 \times 31 = 961$ variables ('8x undersampled')

Total variation denoising





120 linear measurements and $31 \times 31 = 961$ variables ('8x undersampled')

Group lasso

(*e.g.*, Yuan & Lin; Meier, van de Geer, Bühlmann; Jacob, Obozinski, Vert)

- problem:

$$\text{minimize} \quad f(x) + \lambda \sum_{i=1}^{N} \|x_i\|_2$$

  i.e., like lasso, but require groups of variables to be zero or not

- also called $\ell_{1,2}$ mixed norm regularization

- related to **multiple kernel learning** via duality (see Bach et al.)

Joint covariate selection for multi-task learning

(Obozinski, Taskar, Jordan)

- want to fit parameters $x^k \in \mathbf{R}^p$ for each of **multiple** datasets $\mathcal{D}^k$
- either use feature j in all tasks or none of them
- let $x_j = (x_j^1, \ldots, x_j^K)$ for $j = 1, \ldots, p$
- problem:

$$\text{minimize} \quad \sum_{k=1}^K f^k(x^k) + \lambda \sum_{j=1}^p \|x_j\|_2$$

  with variables $x^1, \ldots, x^K \in \mathbf{R}^p$

Structured group lasso

(Jacob, Obozinski, Vert; Bach et al.; Zhao, Rocha, Yu; . . .)

- problem:

$$\text{minimize} \quad f(x) + \sum_{i=1}^{N} \lambda_i \|x_{g_i}\|_2$$

  where $g_i \subseteq [n]$ and $\mathcal{G} = \{g_1, \ldots, g_N\}$

- like group lasso, but the groups can overlap arbitrarily

- particular choices of groups can impose 'structured' sparsity

- *e.g.*, topic models, selecting interaction terms for (graphical) models, tree structure of gene networks, fMRI data

- generalizes to the **composite absolute penalties family**:

$$r(x) = \|(\|x_{g_1}\|_{p_1}, \ldots, \|x_{g_N}\|_{p_N})\|_{p_0}$$

Structured group lasso

(Jacob, Obozinski, Vert; Bach et al.; Zhao, Rocha, Yu; . . .)

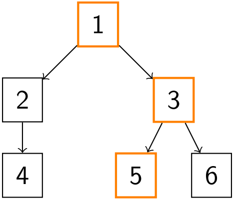
contiguous selection:

- $\mathcal{G} = \{\{1\}, \{5\}, \{1,2\}, \{4,5\}, \{1,2,3\}, \{3,4,5\}, \{1,2,3,4\}, \{2,3,4,5\}\}$
- nonzero variables are contiguous in x, *e.g.*, $x^\star = (0, *, *, 0, 0)$
- can extend the same idea to higher dimensions (*e.g.*, select rectangles)
- *e.g.*, time series, tumor diagnosis, . . .

Structured group lasso

(Jacob, Obozinski, Vert; Bach et al.; Zhao, Rocha, Yu; . . .)

hierarchical selection:



- $\mathcal{G} = \{\{4\}, \{5\}, \{6\}, \{2,4\}, \{3,5,6\}, \{1,2,3,4,5,6\}\}$
- nonzero variables form a rooted and connected subtree
  - if node is selected, so are its ancestors
  - if node is not selected, neither are its descendants

Matrix decomposition

- problem:

$$\begin{array}{ll} \text{minimize} & f_1(X_1) + \cdots + f_N(X_N) \\ \text{subject to} & X_1 + \cdots + X_N = A \end{array}$$

- many choices for the f_i:
  - squared Frobenius norm (least squares)
  - entrywise ℓ_1 norm (sparse matrix)
  - nuclear norm (low rank)
  - sum-$\{$row,column$\}$-norm (group lasso)
  - elementwise constraints (fixed sparsity pattern, nonnegative, ...)
  - semidefinite cone constraint

- easy to solve via ADMM if $\mathbf{prox}_{f_i}$ and $\Pi_{\mathcal{C}}$ are simple enough

Low rank matrix completion

(Candès & Recht; Recht, Fazel, Parrilo)

- problem:

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & X_{ij} = A_{ij}, \quad (i,j) \in \mathcal{D} \end{array}$$

  i.e., find low rank matrix that agrees with observed entries

- *e.g.*, Netflix problem

Robust PCA

(Candès et al.; Chandrasekaran et al.)

- regular PCA is the (nonconvex but solvable) problem

$$\begin{array}{ll} \text{minimize} & \|A - L\|_2 \\ \text{subject to} & \mathbf{rank}(L) \leq k \end{array}$$

i.e., recover rank k matrix L_0 if $A = L_0 + N_0$, where N_0 is noise

- if matrix also has some sparse but large noise, instead solve

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} & L + S = A \end{array}$$

i.e., recover low rank L and sparse corruption S if $A = L_0 + S_0 + N_0$

- sparse + low rank decomposition has other applications (*e.g.*, vision, video segmentation, background subtraction, biology, indexing)

Robust PCA

(Candès et al.; Chandrasekaran et al.)

Structure learning in Gaussian graphical models

(Banerjee et al.; Friedman et al.; Chandrasekaran et al.)

- structure in precision matrix dictates Markov properties of MRF
- learn structure of (observed) Gaussian MRF via ℓ_1 regularized MLE:

$$\begin{array}{ll}\text{minimize} & -l(X;\hat{\Sigma}) + \lambda\|X\|_1 \\ \text{subject to} & X \succeq 0\end{array}$$

- can extend to models with latent variables via

$$\begin{array}{ll}\text{minimize} & -l(S-L;\hat{\Sigma}) + \lambda_1\|L\|_* + \lambda_2\|S\|_1 \\ \text{subject to} & S-L \succeq 0, \quad L \succeq 0\end{array}$$

- many (involved) results on consistency of these estimators

Outline

Proximal operator

(Martinet; Moreau; Rockafellar)

- proximal operator of $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is

$$\mathbf{prox}_{\lambda f}(v) = \underset{x}{\operatorname{argmin}} \left(f(x) + (1/2\lambda)\|x - v\|_2^2 \right)$$

  with parameter $\lambda > 0$

- f may be nonsmooth, have embedded constraints, ...

- **example**: proximal operator of $I_\mathcal{C}$ is $\Pi_\mathcal{C}$

- *many* interpretations

Polyhedra

- projection onto polyhedron $\mathcal{C} = \{x \mid Ax = b,\ Cx \leq d\}$ is a QP

- projection onto affine set $\mathcal{C} = \{x \mid Ax = b\}$ is a linear operator

- box or hyperrectangle $\mathcal{C} = \{x \mid l \preceq x \preceq u\}$:

$$(\Pi_{\mathcal{C}}(v))_k = \begin{cases} l_k & v_k \leq l_k \\ v_k & l_k \leq v_k \leq u_k \\ u_k & v_k \geq u_k, \end{cases}$$

- also simple methods for hyperplanes, halfspaces, simplexes, ...

Norms and norm balls

- **in general**: if $f = \|\cdot\|$ and $\mathcal{B}$ is unit ball of dual norm, then

$$\mathbf{prox}_{\lambda f}(v) = v - \lambda \Pi_{\mathcal{B}}(v/\lambda)$$

- if $f = \|\cdot\|_2$ and $\mathcal{B}$ is the unit ℓ_2 ball, then

$$\Pi_{\mathcal{B}}(v) = \begin{cases} v/\|v\|_2 & \|v\|_2 > 1 \\ v & \|v\|_2 \leq 1 \end{cases}$$

$$\mathbf{prox}_{\lambda f}(v) = \begin{cases} (1 - \lambda/\|v\|_2)v & \|v\|_2 \geq \lambda \\ 0 & \|v\|_2 < \lambda \end{cases}$$

sometimes called 'block soft thresholding' operator

Norms and norm balls

- if $f = \|\cdot\|_1$ and $\mathcal{B}$ is the unit ℓ_∞ ball, then

$$(\Pi_{\mathcal{B}}(v))_i = \begin{cases} 1 & v_i > 1 \\ v_i & |v_i| \leq 1 \\ -1 & v_i < -1 \end{cases}$$

  lets us derive (elementwise) **soft thresholding**

$$\mathbf{prox}_{\lambda f}(v) = (v - \lambda)_+ - (-v - \lambda)_+ = \begin{cases} v_i - \lambda & v_i \geq \lambda \\ 0 & |v_i| \leq \lambda \\ v_i + \lambda & v_i \leq -\lambda \end{cases}$$

- if $f = \|\cdot\|_\infty$ and $\mathcal{B}$ is unit ℓ_1 ball, simple algorithms available

Matrix functions

- suppose convex $F : \mathbf{R}^{m\times n} \to \mathbf{R}$ is orthogonally invariant:

$$F(QX\tilde{Q}) = F(X)$$

  for all orthogonal Q, $\tilde{Q}$

- then $F = f \circ \sigma$ and

$$\mathbf{prox}_{\lambda F}(A) = U\,\mathbf{diag}(\mathbf{prox}_{\lambda f}(d))V^T$$

  where $A = U\,\mathbf{diag}(d)V^T$ is the SVD of A and $\sigma(A) = d$

- *e.g.*, $F = \|\cdot\|_*$ has $f = \|\cdot\|_1$ so $\mathbf{prox}_{\lambda F}$ is 'singular value thresholding'

Proximal gradient method

(*e.g.*, Levitin & Polyak; Mercier; Chen & Rockafellar; Combettes; Tseng)

- problem form

$$\text{minimize} \quad f(x) + g(x)$$

  where f is smooth and $g : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is closed proper convex

- method:

$$x^{k+1} := \mathbf{prox}_{\lambda^k g}(x^k - \lambda^k \nabla f(x^k))$$

- special case: projected gradient method (take $g = I_\mathcal{C}$)

Accelerated proximal gradient method

(Nesterov; Beck & Teboulle; Tseng)

- problem form

$$\text{minimize} \quad f(x) + g(x)$$

  where f is smooth and $g : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is closed proper convex

- method:

$$\begin{aligned} y^{k+1} &:= x^k + \omega^k \left(x^k - x^{k-1}\right) \\ x^{k+1} &:= \mathbf{prox}_{\lambda^k g}\left(y^{k+1} - \lambda^k \nabla f(y^{k+1})\right) \end{aligned}$$

  works for, *e.g.*, $\omega^k = k/(k+3)$ and particular λ^k

- faster in both theory and practice

ADMM

(*e.g.*, Gabay & Mercier; Glowinski & Marrocco; Boyd et al.)

- problem form

$$\text{minimize} \quad f(x) + g(x)$$

  where $f,\ g : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ are closed proper convex

- method:

$$\begin{aligned} x^{k+1} &:= \mathbf{prox}_{\lambda f}(z^k - u^k) \\ z^{k+1} &:= \mathbf{prox}_{\lambda g}(x^{k+1} + u^k) \\ u^{k+1} &:= u^k + x^{k+1} - z^{k+1} \end{aligned}$$

- basically, always works

Examples

- (accelerated) proximal gradient for elastic net:

  ❶ gradient step for smooth loss (*e.g.*, logistic, least squares, ...)
  ❷ shrinkage and elementwise soft thresholding

- ADMM for multi-task learning with joint covariate selection:

  ❶ evaluate $\mathbf{prox}_{f^k}$ (in parallel for each dataset)
  ❷ block soft thresholding (in parallel for each feature)
  ❸ dual update

- ADMM for robust PCA:

  ❶ singular value thresholding
  ❷ elementwise soft thresholding
  ❸ dual update

Outline

Conclusions

questions? (for details, see the papers)

some review papers:

Bach et al. *Optimization with sparsity-inducing penalties*. FTML, 2011.

Boyd. ℓ_1 *methods for convex-cardinality problems*. EE 364b Notes.

Boyd et al. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. FTML, 2011.

Bruckstein et al. *From sparse solutions of systems of equations to sparse modeling of signals and images*. SIAM Review, 2009.

Hastie et al. *The Elements of Statistical Learning*, chapter 18 (high-dimensional problems).

Parikh and Boyd. *Proximal algorithms*. To appear in FTOC, 2013.