

# Gradient Algorithms for Regularized Optimization

Stephen Wright

University of Wisconsin-Madison

SPARS11, Edinburgh, June 2011

Partial overview of some techniques from computational optimization of possible relevance to sparse reconstruction.

- Illustrating the effectiveness of  $\ell_1$  for sparsity.
- Other regularizers for other structures.
- Prox-Linear methods
  - Implementation for different regularizers
  - Extensions
  - Enhancements
  - Identifying the optimal manifold and using higher-order info
- Augmented Lagrangian
- Stochastic gradient methods: some recent results.

# Sparse / Regularized Optimization

Many applications need **structured, approximate** solutions of optimization formulations, rather than exact solutions.

- More Useful, More Credible
  - Structured solutions are easier to comprehend / use / actuate.
  - They correspond better to prior knowledge.
  - Extract just essential meaning from the data, not less important effects.
- Less Data Needed
  - Structured solution lies in lower-dimensional spaces than ambient space  
⇒ need to gather / sample less data to capture it.

The structural requirements have deep implications for how we formulate and solve these problems. There's a lot of variety in the properties and contexts of the various applications.

Compressed sensing problems fit these principles. Also machine learning and many application areas e.g. medical imaging, geophysics, power grids, control.

# $\ell_1$ and Sparsity

Use of  $\|x\|_1$  has long been known to promote sparsity in  $x$ . Also, it's convex, and avoids discrete variables (associated with limits on cardinality  $\|\cdot\|_0$ ) in the formulation.

**Weighted form:**  $\min f(x) + \tau\|x\|_1$ , for some  $\tau > 0$ .

**$\ell_1$ -constrained form** (variable selection):  $\min f(x)$  subject to  $\|x\|_1 \leq T$ .

**Function-constrained form:**  $\min \|x\|_1$  subject to  $f(x) \leq \bar{f}$ .

In compressed sensing,  $\|\cdot\|_1$  is (provably) a perfect surrogate for  $\|\cdot\|_0$ .

*Recover  $x \in \mathbb{R}^n$  from observations  $y \in \mathbb{R}^m$  given  $Ax = y$  with known sensing matrix  $A \in \mathbb{R}^{m \times n}$ .*

*Additionally, **know that  $x$  is sparse:**  $\|x\|_0 \ll n$ .*

# When Does $\ell_1$ Work?

Elementary analysis from W. Yin and Y. Zhang, *SIAG Views and News* 19 (2008), using Kashin (1977) and Garnaev and Gluskin (1984).

Suppose that  $\bar{x}$  is the *minimum-cardinality solution* of the underdetermined linear equations  $Ax = y$ , where  $A \in \mathbb{R}^{m \times n}$  with  $m < n$ .

$$\bar{x} = \arg \min \|x\|_0 \text{ s.t. } Ax = y.$$

- $S \subset \{1, 2, \dots, n\}$  be the support of  $\bar{x}$ ;
- $k := \|\bar{x}\|_0 = |S|$ ;
- $Z = S^c$ .

The 1-norm form is:

$$\min \|x\|_1 \text{ s.t. } Ax = y. \quad (1)$$

$\bar{x}$  solves this problem too provided

$$\|\bar{x} + v\|_1 \geq \|\bar{x}\|_1 \text{ for all } v \in N(A).$$

$$\begin{aligned}
\|\bar{x} + v\|_1 &= \|\bar{x}_S + v_S\|_1 + \|v_Z\|_1 \\
&\geq \|\bar{x}_S\|_1 + \|v_Z\|_1 - \|v_S\|_1 \\
&= \|\bar{x}\|_1 + \|v\|_1 - 2\|v_S\|_1 \\
&\geq \|\bar{x}\|_1 + \|v\|_1 - 2\sqrt{k}\|v\|_2.
\end{aligned}$$

Hence,  $\bar{x}$  solves (1) provided that

$$\frac{1}{2} \frac{\|v\|_1}{\|v\|_2} \geq \sqrt{k} \text{ for all } v \in N(A).$$

In general we have only:

$$1 \leq \frac{\|v\|_1}{\|v\|_2} \leq \sqrt{n}$$

However this ratio tends to be **significantly larger than 1** if  $v$  is restricted to a random subspace.

Specifically, if the elements of  $A \in \mathbb{R}^{m \times n}$  are chosen iid from  $N(0, 1)$ , we have with high probability that

$$\frac{\|v\|_1}{\|v\|_2} \geq \frac{C\sqrt{m}}{\sqrt{\log(n/m)}}, \text{ for all } v \in N(A),$$

for some constant  $C$ . (Concentration of measure.)

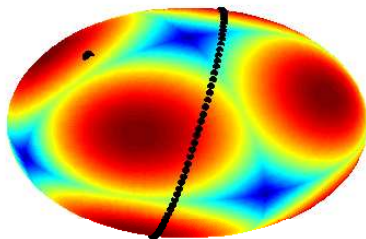
Thus, with high prob,  $\bar{x}$  solves (1) if

$$m \geq \frac{4}{C^2} k \log n.$$

Conclusion: The required number  $m$  of random linear observations is a multiple of  $k \log n$  — typically much less than  $n$ .

# Ratio $\|v\|_1/\|v\|_2$ in $\mathbb{R}^3$

Plotting  $\|v\|_1$  on sphere  $\{v : \|v\|_2 = 1\}$ . Blue:  $\|v\|_1 \approx 1$ . Red:  $\|v\|_1 \approx \sqrt{3}$ . Ratio is smallest along the principal axes.

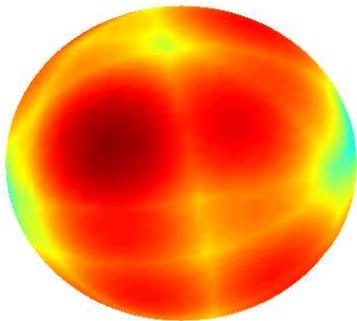


Dot:  $N(A)$  for a random  $A \in \mathbb{R}^{2 \times 3}$ .  
Equator:  $N(A)$  for a random  $A \in \mathbb{R}^{1 \times 3}$ .  
(Both usually avoid smaller values of  $\|v\|_1$ .)



# Ratio $\|v\|_1/\|v\|_2$ on Random Null Spaces

Random  $A \in \mathbb{R}^{4 \times 7}$ , showing ratio  $\|v\|_1$  for  $v \in N(A)$  with  $\|v\|_2 = 1$



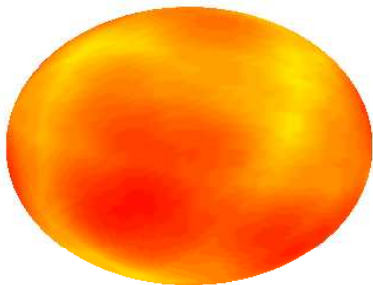
Blue:  $\|v\|_1 \approx 1$ . Red: ratio  $\approx \sqrt{7}$ .

$\|v\|_1$  is well away from the lower bound of 1 over the whole nullspace.

# Ratio $\|v\|_1/\|v\|_2$ on Random Null Spaces

The effect grows more pronounced as  $m/n$  grows.

Random  $A \in \mathbb{R}^{17 \times 20}$ , showing ratio  $\|v\|_1$  for  $v \in N(A)$  with  $\|v\|_2 = 1$ .



Blue:  $\|v\|_1 \approx 1$ . Red:  $\|v\|_1 \approx \sqrt{20}$ .

$\|v\|_1$  is far from lower bound throughout.

# Extensions of $\ell_1$ : Group Regularizers

There may be a natural relationship between some components of  $x$ . We could thus group the components, and **select or deselect at the group level**.

Use “sum of  $\ell_\infty$ ” or “sum of  $\ell_2$ ” regularizers:

$$\sum_{k=1}^m \|x_{[k]}\|_\infty, \quad \sum_{k=1}^m \|x_{[k]}\|_2,$$

where  $[k]$  (for  $k = 1, 2, \dots, m$ ) represent subsets of the components of  $x$ .

The subvectors  $x_{[k]}$  can be **overlapping** or **non-overlapping**. (The latter are generally easier to deal with.)

# Examples: Separable Groups

Simultaneous variable selection (select a subset of variables to explain a number of observation vectors simultaneously, for a fixed design matrix) (e.g. Turlach, Venables, Wright, 2005):

$$\min_X \frac{1}{2} \|Y - AX\|_F^2 + \tau \sum_{i=1}^m \|X_{i,\cdot}\|_\infty.$$

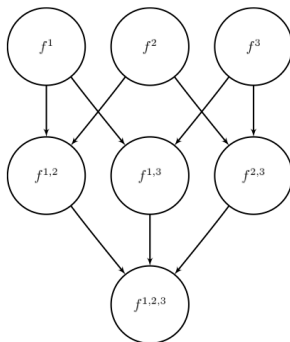
Fitting observations sparsely from a fixed dictionary:

$$\min_X \frac{1}{2} \|Y - AX\|_F^2 + \tau \sum_{j=1}^n c(X_{\cdot,j}),$$

where  $c(\cdot) = \|\cdot\|_\infty$ , or a more general function. e.g. (Jenatton et al, 2010) define  $c$  to be a regularizer for hierarchical overlapping groups, and minimizes over  $A$  as well (sparse dictionary learning).

# Examples: Overlapping Groups

(Ding et al, 2011): Learning graphical models from multivariate Bernoulli outcomes. Each group = all descendants of a node in a directed graph.



# Examples: Tree-Structured Groups

In a wavelet representation  $U = Wx$  coefficients  $x$  can be arranged in a quadtree, exposing a hierarchical relationship between them.



Subtree structure of natural images can be exposed by

- imposing it explicitly (Baraniuk et al., 2010);
- inducing via group regularizers (Jenatton et al., 2010: For any two non-disjoint groups  $[k]$  and  $[l]$ , have  $[k] \subset [l]$  or  $[l] \subset [k]$ ) Also (Rao et al., 2011; talk on Tuesday).

# Reformulating Overlapping Groups

Formulate the problem with overlapping-group regularizer as one with separable groups plus equality constraints, using **replication** of variables:

$$\min_x f(x) + \tau \sum_{k=1}^m \|x_{[k]}\|$$

can be rewritten as

$$\min_{\bar{x}, x^1, x^2, \dots, x^m} f(\bar{x}) + \tau \sum_{k=1}^m \|x_{[k]}^k\| \quad \text{s.t. } x_i^k = \bar{x}_i, \quad i \in [k], \quad k = 1, 2, \dots, m.$$

By using a quadratic penalty to (approximately) enforce the constraints, can get an approximate solution to the overlapping-group regularized problem by solving a non-overlapping-group problem:

$$\min_{\bar{x}, x^1, x^2, \dots, x^m} f(\bar{x}) + \tau \sum_{k=1}^m \|x_{[k]}^k\| + \frac{\mu}{2} \sum_{k=1}^m \sum_{i \in [k]} (x_i^k - \bar{x}_i)^2.$$

# An Alternative Overlapping Group Norm

Can get an alternative overlapping norm by relaxing the constraints  $\bar{x}_i = x_i^k$ ,  $k = 1, 2, \dots, m$  in the previous formulation to

$$\bar{x}_i = \frac{1}{N_i} \sum_{k: i \in [k]} x_i^k,$$

where  $N_i$  is the number of groups containing element  $i$ . By substituting for  $\bar{x}$ , we again obtain a formulation with non-overlapping-groups, but the solutions have different properties.

For this regularizer, the support of optimal  $x$  tends to be a **union of selected groups**, whereas for the usual group regularizer the support tends to be the **complement of the union of non-selected groups**.

(Jacob, Obozinski, Vert, 2009)



# Total-Variation Regularization

Given intensity measures  $U_{ij}$  for  $i, j = 1, 2, \dots, N$  (a 2D grid), define the *variation* at grid point  $(i, j)$  as

$$\left\| \begin{bmatrix} U_{i+1,j} - U_{ij} \\ U_{i,j+1} - U_{ij} \end{bmatrix} \right\|_2.$$

(zero iff  $U_{ij}$ ,  $U_{i+1,j}$ ,  $U_{i,j+1}$  all have the same intensity).

**Total Variation** obtained by summing across the grid:

$$\text{TV}(U) := \frac{1}{N^2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \left\| \begin{bmatrix} U_{i+1,j} - U_{ij} \\ U_{i,j+1} - U_{ij} \end{bmatrix} \right\|_2.$$

Forces most grid points  $(i, j)$  to have the same intensities as their neighbors. (Rudin, Osher, Fatemi, 1992)

**Denoising:** Given observed intensities  $F \in \mathbb{R}^{N \times N}$ , solve

$$\min_{U \in \mathbb{R}^{N \times N}} \frac{1}{2} \|U - F\|_F^2 + \tau \text{TV}(U).$$

# Matrix Estimation / Completion

Given an  $m \times n$  matrix  $M$  in which only certain elements are known:

$$\Omega \subset \{(i, j) \mid i = 1, 2, \dots, m, \ j = 1, 2, \dots, n\}.$$

Find a matrix  $X$  with “nice structure” such that  $X_{ij} \approx M_{ij}$  for  $(i, j) \in \Omega$ .

Example: **Netflix**

Desirable structures:

- Low rank: induced by nuclear norm  $\|X\|_*$  (sum of singular values) (Recht, Fazel, Parrilo, 2010) or “max norm” (Lee et al., 2010).
- Sparsity: Induced by element-wise 1-norm:  $\sum_{i,j} |X_{ij}|$ .
- Both: combine these regularizers.

# Algorithms: Many Techniques Used

- Large-scale optimization: optimal first-order, gradient projection, second-order, continuation, coordinate relaxation, interior-point, augmented Lagrangian, conjugate gradient, semismooth Newton ...
- Nonsmooth optimization: cutting planes, subgradient methods, successive approximation, smoothing, prox-linear methods, ...
- Dual and primal-dual formulations / methods
- Numerical linear algebra
- Stochastic approximation, sampled-average approximation.
- Heuristics

Also a LOT of domain-specific knowledge about the problem structure and the type of solution demanded by the application.

# Prox-Linear Methods: Foundation

For the setting

$$\min_x \phi_\tau(x) := f(x) + \tau c(x).$$

At  $x^k$ , solve this subproblem for new iterate  $x^{k+1}$ :

$$\text{PLS: } x^{k+1} = \arg \min_z \nabla f(x^k)^T (z - x^k) + \tau c(z) + \frac{1}{2\alpha_k} \|z - x^k\|_2^2,$$

for some choice of  $\alpha_k > 0$  (see below).

Works well when this subproblem is easy to formulate and solve.

If  $c$  is vacuous, this reduces to gradient descent, with a line search.

A fundamental approach that goes by different names in different settings, e.g. IST, Forward-Backward Splitting, SpaRSA.

# The Prox-Linear Subproblem

Requires evaluation of  $\nabla f$ .

- In compressed sensing, requires multiplication by  $A$  and  $A^T$  — inexpensive for partial FFT, wavelets, etc.
- Variants use subvector of  $\nabla f$  or estimate of  $\nabla f$  based on only part of the data. (See below.)

Formulate the subproblem equivalently as

$$\min_z \frac{1}{2} \left\| z - \left[ x^k - \alpha_k \nabla f(x^k) \right] \right\|_2^2 + \tau \alpha_k c(z),$$

which is an application of the Moreau proximity operator (“shrink operator”) associated with  $c$ :

$$S_\sigma(x) = \arg \min_z \frac{1}{2} \|z - x\|_2^2 + \sigma c(z).$$

# Calculating the Shrink Operator

There are closed-form solutions for some important regularizers:

$$c(x) = \|x\|_1 : \quad S_\sigma(x)_i = \begin{cases} 0 & \text{if } x_i \in [-\sigma, \sigma], \\ x_i + \sigma & \text{if } x_i < -\sigma, \\ x_i - \sigma & \text{if } x_i > \sigma. \end{cases}$$

$$c(x) = \|x\|_2 : \quad \begin{cases} 0 & \text{if } \|x\|_2 \leq \sigma; \\ (1 - \sigma/\|x\|_2)x & \text{otherwise.} \end{cases}$$

$c(x) = \|x\|_\infty$  : closed-form solution obtained after a sort of  $|x_i|$ ,  $i = 1, 2, \dots$ ,

Separable groups: Separate the shrink operator and solve separately:

$$S_\sigma(x_{[k]}) = \arg \min_{z_{[k]}} \frac{1}{2} \|z_{[k]} - x_{[k]}\|_2^2 + \sigma c(z_{[k]}).$$

# Tree-structured Groups

For tree-structure groups, the sum-of- $\ell_2$  and sum-of- $\ell_\infty$  can also be calculated efficiently (Jenatton et al, 2010).

- Order the groups such that either  $[k] \subset [l]$  or  $[k] \cap [l] = \emptyset$  for all  $k < l$ .
- Perform the shrink in sequence to subsets  $[1], [2], \dots, [m]$ .

One pass through the groups suffices.

# (General) overlapping groups

Dual formulation of  $S_\sigma(x)$ :

$$\max_{\xi_1, \xi_2, \dots, \xi_m} -\frac{1}{2} \|x - \sum_{k=1}^m \xi_k\|_2^2 + \frac{1}{2} \|x\|_2^2 \quad \text{s.t.} \quad \|\xi_k\|_* \leq \sigma, \quad (\xi_k)_i = 0 \text{ for all } i \notin [k].$$

where  $\|\cdot\|_*$  is the dual of the norm in the regularized (i.e.  $\|\cdot\|_2$  for  $\|\cdot\|_2$ ,  $\|\cdot\|_1$  for  $\|\cdot\|_\infty$ ).

This is a convex quadratic program. Can solve by

- block coordinate relaxation
- gradient projection
- algorithm for quadratic min-cost network flow (Mairal et al, 2010).



# Shrinking with the TV-norm

$$\begin{aligned} \min_{U \in \mathbb{R}^{N \times N}} \quad & \frac{1}{2} \|U - F\|_F^2 + \tau \text{TV}(U) \\ = \quad & \frac{1}{2} \|U - F\|_F^2 + \frac{\tau}{N^2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \|A_{ij} U\|_2, \end{aligned}$$

where  $A_{ij}$  is an  $2 \times N^2$  matrix with four  $\pm 1$ s and the rest zeros. **Dual** is

$$\max_{w \in W} \quad \frac{1}{2} \|Aw - \text{vec}(F)\|_2^2,$$

where  $W = \{(w_1, w_2, \dots, w_{N^2}) \mid w_i \in \mathbb{R}^2, \|w_i\|_2 \leq 1\}$ .  $A$  is an  $N^2 \times 2N^2$  matrix whose columns are  $A_{ij}^T$ .

Can solve dual efficiently by gradient projection (Zhu, Wright, Chan, 2010) with various choices of steplength.

Many other algorithms, e.g. (Chambolle, 2004), primal-dual method of (Zhu, Chan, 2008)

# Shrinking with the Nuclear Norm

Shrink operator with the nuclear norm is

$$\min_Z \frac{1}{2} \|Z - Y^k\|_F^2 + \sigma \|Z\|_*.$$

Can solve explicitly using a singular value decomposition of  $Y^k$ , followed by an adjustment (shrink) of the singular values.

**Code:** SVT: Compute a partial SVD (largest singular values) using Lanczos. (Cai, Candès, Shen, 2008)

Shrinking with the max-norm is also efficient (Lee et al. 2010).

# A Prox-Linear Method

e.g. one of the variants of SpaRSA: (Wright, Nowak, Figueiredo, 2008).

At iteration  $k$ :

- Solve for current  $\alpha_k$  to find candidate solution  $x^{k+}$ :

$$x^{k+} = \arg \min_z \nabla f(x^k)^T (z - x^k) + \tau c(z) + \frac{1}{2\alpha_k} \|z - x^k\|_2^2,$$

- Decrease  $\alpha_k$  as needed until sufficient decrease is obtained:

$$\phi_\tau(x^k) - \phi_\tau(x^{k+}) \geq \|x^{k+} - x^k\|_2^3.$$

- Increase  $\alpha_k$  by a constant factor (but enforce  $\alpha_k \leq \alpha_{\max}$ ) in preparation for next iteration.

All accumulation points are minimizers. (Not surprising, as it's a convex problem.) No global rate in general (i.e. linear/exponential convergence or sublinear e.g.  $1/k$  rate).

# Many Variants / Enhancements

- Nonmonotone method using a Barzilai-Borwein choice of parameter  $\alpha_k$  (another SpaRSA variant).
- Basic IST: Chooses  $\alpha_k \equiv \bar{\alpha} < 1/L$ . Guarantees descent in  $\phi_\tau$  at every iteration.
- Continuation in the regularization parameter  $\tau$ . Solve a sequence of problems for different  $\tau$ , from large to small, and warm start.
- Block Coordinate Relaxation: Calculate just a partial gradient (subvector of  $\nabla f$ ) at each iteration.
- Accelerated first-order methods (e.g. FISTA, NESTA).
- Debiasing: When  $c(x) = \|\cdot\|_1$ , switch to a local “debiasing” phase once the correct set of nonzeros is identified and discard the regularization term. RIP  $\Rightarrow$  linear convergence in this phase.

# Accelerated First-Order Methods

Can exploit the research on methods for smooth convex optimization that use gradients, but do better than simply stepping in the negative gradient direction  $-\nabla f(x)$ . (Nesterov)

They generate two (or three) intertwined sequences. Typically:

- Get the next  $x$ -sequence iterate from a short gradient-descent step from the latest  $y$ -sequence element
- Get the next  $y$ -sequence element by extrapolating from the last two  $x$ -sequence iterates.

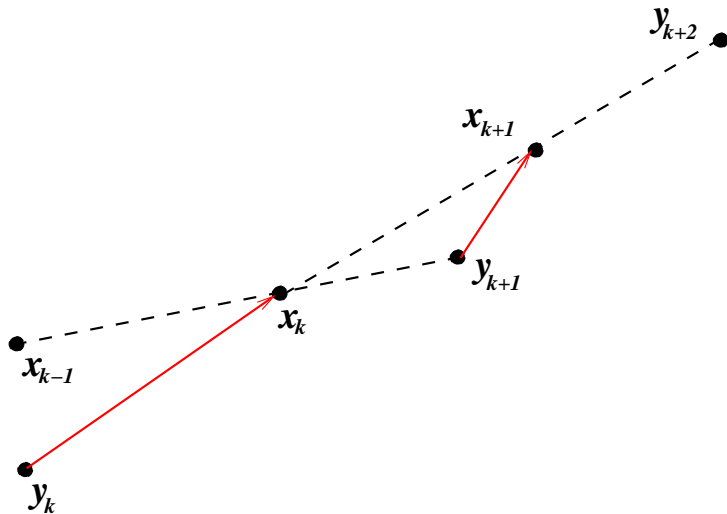
FISTA (Beck and Teboulle, 2008):  $\min f$ ,  $L = \text{Lipschitz const for } \nabla f$ :

0: Choose  $x_0$ ; set  $y_1 = x_0$ ,  $t_1 = 1$ ;

$k$ :  $x_k \leftarrow y_k - \frac{1}{L} \nabla f(y_k)$ ;

$$t_{k+1} \leftarrow \frac{1}{2} \left( 1 + \sqrt{1 + 4t_k^2} \right);$$
$$y_{k+1} \leftarrow x_k + \frac{t_k - 1}{t_{k+1}} (x_k - x_{k-1}).$$

# Two Sequences: $\{x_k\}$ and $\{y_k\}$



Analysis is short but not very intuitive.  $f(x_k)$  converges to its optimal value  $f^*$  at a “fast” sublinear rate of  $O(1/k^2)$ .

Can extend to the regularized problem  $\min f(x) + \tau c(x)$  by replacing the step  $y_k \rightarrow x_k$  by the prox-linear subproblem, with  $\alpha_k \equiv 1/L$ .

Practically, less sensitive to  $\tau$  than prox-linear.

Similar approaches can achieve a linear / exponential rate when  $f$  is *strongly* convex. (Nesterov, 2004)

# Block-Coordinate Relaxation

Suitable for problems with separable group regularizers.

$$\min_x f(x) + \tau \sum_{l=1}^m c_l(x_{[l]}).$$

where  $f$  is smooth; each  $P_q$  is closed, proper, convex. Assume that  $\{x_{[q]} : q \in Q\}$  is a partition of the components of  $x$ .

At iteration  $x^k$ , choose a subset  $Q_k \subset \{1, 2, \dots, m\}$  and solve  $\mu_k \in [\mu_{\min}, \mu_{\text{top}}]$ :

$$\min_d \nabla f(x^k)^T(z - x^k) + \frac{\mu_k}{2} \|z - x^k\|_2^2 + \tau \sum_{l \in Q_k} c_l(z_{[l]}) \quad \text{s.t. } z_{[l]} = x_{[l]}^k \text{ for } l \notin Q_k.$$

Increase  $\mu_k$  until sufficient decrease, e.g.

$$\phi_\tau(x^k) - \phi_\tau(z) \geq \|z - x^k\|_2^3.$$

Set  $x^{k+1} = z$ .



“Generalized Gauss-Seidel” condition: For some  $T$  and all  $k$ , require

$$Q_k \cup Q_{k-1} \cup \cdots \cup Q_{k-T} = \{1, 2, \dots, m\}.$$

Possibly replace  $x^{k+1}$  by an improved step e.g. by improving  $\phi_\tau$  further on the current manifold defined by  $Q_k$  and  $z^k$ , possibly using second-order information.

**Global Convergence Result:** If  $\nabla f$  is locally Lipschitz on a neighborhood of the level set  $\{x \mid \phi_\tau(x) \leq \phi_\tau(x^0)\}$  and if  $\{\phi_\tau(x^k)\}$  is bounded below, then all accumulation points are critical.

(Tseng, Yun, 2009; Wright, 2010).

# Prox-Linear for a Composite Minimization Framework

(Lewis, Wright, 2008) Analyze convergence of a basic descent algorithm for prox-linear by embedding in the framework of **composite minimization**:

$$\min h(p(x))$$

where  $p : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is smooth,  $h : \mathbb{R}^m \rightarrow \mathbb{R}$  is partly smooth and *prox-regular* (i.e. locally convex to within a quadratic fudge term).

- Allows some nonconvex regularizers to be used, e.g.

$$|x|_* = \sum_{i=1}^n (1 - e^{-\alpha|x|_i}),$$

for some  $\alpha > 0$ . (Mangasarian, 1999), (Jokar and Pfetsch, 2007).

- $h$  can be extended-valued — can enforce hard constraints.

Subproblems are:

$$\min_z h(p(x^k) + \nabla p(x^k)^T(z - x^k)) + \frac{1}{2\alpha_k} \|z - x^k\|_2^2.$$

The prox-linear framework is broadly the same in this setting, but we may need to modify the  $z$  obtained from the subproblem slightly to ensure that

$$p(x^k) + \nabla p(x^k)^T(z - x^k) \in \text{dom } h;$$

Put  $\phi_\tau = f + \tau c$  into this framework by defining

$$p(x) := \begin{bmatrix} f \\ x \end{bmatrix}, \quad h(p) := p_1 + \tau c([p_2, p_3, \dots, p_{n+1}]).$$

# Manifolds and Partial Smoothness

Although the regularizer  $c$  is usually nonsmooth, we can often identify a smooth manifold in  $\mathbb{R}^n$  along which  $c$  behaves like a smooth function.

**Manifold:** Surface in  $\mathcal{M} \subset \mathbb{R}^n$  that can be parametrized by smooth vector functions in the neighborhood of some point  $\bar{x} \in \mathcal{M}$ , e.g.  $z(s) \in \mathcal{M}$  where  $z : \mathbb{R}^t \rightarrow \mathbb{R}^n$ , for all  $s$  in a neighborhood of  $0 \in \mathbb{R}^t$ , where  $z$  is smooth near 0.

**Partial Smoothness:** (Lewis, 2003)  $\phi$  is partly smooth with respect to manifold  $\mathcal{M}$  at a point  $\bar{x} \in \mathcal{M}$  if it behaves smoothly along  $\mathcal{M}$ , with no collapse in dimension of its generalized gradient.

# Manifolds and Partial Smoothness

Although the regularizer  $c$  is usually nonsmooth, we can often identify a smooth manifold in  $\mathbb{R}^n$  along which  $c$  behaves like a smooth function.

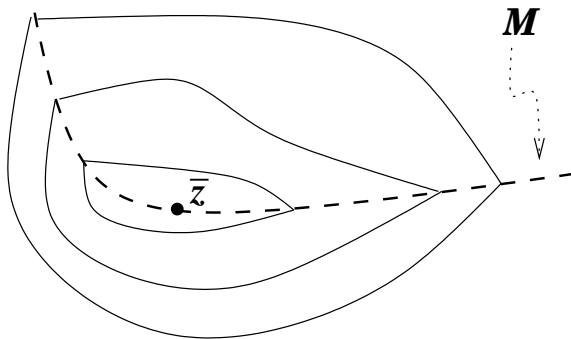
**Manifold:** Surface in  $\mathcal{M} \subset \mathbb{R}^n$  that can be parametrized by smooth vector functions in the neighborhood of some point  $\bar{x} \in \mathcal{M}$ , e.g.  $z(s) \in \mathcal{M}$  where  $z : \mathbb{R}^t \rightarrow \mathbb{R}^n$ , for all  $s$  in a neighborhood of  $0 \in \mathbb{R}^t$ , where  $z$  is smooth near 0.

**Partial Smoothness:** (Lewis, 2003)  $\phi$  is partly smooth with respect to manifold  $\mathcal{M}$  at a point  $\bar{x} \in \mathcal{M}$  if it behaves smoothly along  $\mathcal{M}$ , with no collapse in dimension of its generalized gradient.

- (i)  $\phi|_{\mathcal{M}}$  is  $C^2$ ;
- (ii)  $\phi$  is subdifferentially regular at all  $z \in \mathcal{M}$  near  $\bar{z}$ , with  $\partial h(z) \neq \emptyset$ ;
- (iii)  $\text{aff } \partial h(\bar{z})$  is a translate of  $N_{\mathcal{M}}(\bar{z})$ ;
- (iv)  $\partial h : \mathcal{M} \rightrightarrows \mathbb{R}^m$  is continuous at  $\bar{z}$ .

# Examples: Partial Smoothness

Contours of a function on  $\mathbb{R}^2$  partly smooth at  $\bar{z}$  with respect to the one-dimensional manifold  $\mathcal{M}$ .



**Key Example:** If  $c(x) = \|\cdot\|_1$ , the manifold  $\mathcal{M}$  at  $\bar{x}$  is the set of points  $x$  near  $\bar{x}$  with the **same nonzero structure** as  $\bar{x}$ .

# Identification of $\mathcal{M}$

If the optimum  $x^*$  lies on a manifold  $\mathcal{M}$ , it may be possible to identify  $\mathcal{M}$  without knowing  $x^*$ . Thus, we could apply fast methods to the restriction  $\phi|_{\mathcal{M}}$  during final steps — possibly much lower dimension than the full space  $\mathbb{R}^n$ .

Requires a nondegeneracy condition: Replace

$$\text{criticality: } 0 \in \partial\phi_\tau(x^*)$$

$$\text{by } \textit{strict} \text{ criticality: } 0 \in \text{ri } \partial\phi_\tau(x^*).$$

For  $\phi_\tau(x) = f(x) + \tau\|x\|_1$ , the criticality conditions are:

$$[\nabla f(x^*)]_i \begin{cases} = \tau & \text{if } x_i^* < 0, \\ = -\tau & \text{if } x_i^* > 0, \\ \in [-\tau, \tau] & \text{if } x_i^* = 0, \end{cases}$$

whereas strict criticality replaces  $[-\tau, \tau]$  by the open interval  $(-\tau, \tau)$  in the last line (i.e. no “borderline” components).

# Identification of $\mathcal{M}$

**Identification Result:** If the Prox-Linear algorithm converges to a nondegenerate point  $x^*$ , if  $f$  is locally Lipschitz there, and  $c$  is partly smooth at  $x^*$  with respect to manifold  $\mathcal{M}$ , then  $x^k \in \mathcal{M}$  for all  $k$  sufficiently large.

For the  $\ell_1$  case, this means that for all  $k$  sufficiently large,  $x^k$  has the same nonzero structure as  $x^*$ .

Can modify prox-linear algorithms to seek an enhancement of each new iterate, i.e. after calculating an  $x^{k+1}$  giving sufficient decrease from the prox-linear subproblem, possibly replace it by an enhanced point that

- Decreases the objective further;
- Is not too much further away from  $x^k$ ;
- Lies on the same manifold as the original  $x^{k+1}$ .



# Reduced Second-Order Steps

Use curvature information to enhance steps on the reduced problem  $\phi_\tau|_{\mathcal{M}}$ .

This has been done for *regularized logistic regression*:  $f(x) + \tau\|x\|_1$ , where  $f$  is a log-likelihood function of the form:

$$f(x) = \sum_{i=1}^m \ell(d_i, y_i; x).$$

Abuse notation:  $\mathcal{M} \subset \{1, 2, \dots, n\}$  are the likely nonzeros of  $x^*$  (determined heuristically, but supported by the identification theory above). Enhance  $x^k$  by solving for the reduced Newton direction:

$$[\nabla^2 f(x^k)]_{\mathcal{M}\mathcal{M}} \delta = -[\nabla f(x^k)]_{\mathcal{M}}.$$

(Shi et al, 2008). Can approximate the Hessian by sampling terms randomly from the summation over  $i = 1, 2, \dots, m$  (Byrd et al, 2010).

# Augmented Lagrangian

A classical method: For closed convex  $\Omega$ :

$$\min_{x \in \Omega} p(x) \text{ subject to } Ax = b.$$

Generate iterates  $x^k$  together with Lagrange multiplier estimates  $\lambda^k$  from:

- $x^k$  is approximate solution of

$$\min_{x \in \Omega} p(x) + (\lambda^k)^T (Ax - b) + \frac{\mu_k}{2} \|Ax - b\|_2^2;$$

- update Lagrange multipliers:

$$\lambda^{k+1} = \lambda^k + \mu_k (Ax^k - b).$$

(If  $p$  convex, need only  $\mu_k > 0$ .)

# Constrained Formulation

Given

$$\min_x f(x) + \tau c(x),$$

“duplicate” the variable and write as an equality constrained problem:

$$\min_{z,u} f(z) + \tau c(u) \text{ subject to } u = z.$$

Augmented Lagrangian:

$$(z^k, u^k) := \min_{z,u} f(z) + \tau c(u) + (\lambda^k)^T (u - z) + \frac{\mu_k}{2} \|u - z\|_2^2,$$
$$\lambda^{k+1} := \lambda^k + \mu_k (u^k - z^k).$$

The  $\min_{z,u}$  problem is usually still too hard to solve ( $u$  and  $z$  are coupled via final penalty term). However can take **alternating steps** in  $z$  and  $u$ .

# Alternating Directions

(Eckstein and Bertsekas, 1992)

$$z^k := \min_z f(z) + \tau c(u^{k-1}) + (\lambda^k)^T (u^{k-1} - z) + \frac{\mu_k}{2} \|u^{k-1} - z\|_2^2,$$

$$u^k := \min_u f(z^k) + \tau c(u) + (\lambda^k)^T (u - z^k) + \frac{\mu_k}{2} \|u - z^k\|_2^2,$$

$$\lambda^{k+1} := \lambda^k + \mu_k (u^k - z^k).$$

Approximate minimization for  $z$  and  $u$  may now be much simpler. e.g. for compressed sensing:

- The second minimization is the “shrink operator” (easy);
- The first one is linear system with coefficient matrix  $(A^T A + \sigma I)$  (solve approximately - can be efficient in special cases).

The subproblems are **not vastly different from prox-linear subproblems**:

- $\lambda^k$  is asymptotically similar to the gradient term in prox-linear;
- The quadratic term is the same.

# Use in Sparse Optimization

Extensions and variants of these ideas have been much studied recently, and applied in various contexts. Examples:

- Compressed sensing (Yang and Zhang, 2009; **Code:** YALL1), (Goldfarb, Ma, Scheinberg, 2010)
- Image processing (Figueiredo and Bioucas-Dias, 2010, 2011; Goldstein and Osher, 2008)
- Video processing, matrix completion, sparse principal components (Goldfarb, Ma, Scheinberg, 2010).

Can be melded with (accelerated) first-order methods (Goldfarb, Ma, Scheinberg, 2010).

# Stochastic Gradient Algorithms

Typically work with objective  $f$  for which:

- $f$  convex but possibly nonsmooth.
- Can't get function values  $f(x)$  cheaply.
- At any feasible  $x$ , have access only to an unbiased estimate of the subgradient  $\partial f$ .

Common settings are:

$$f(x) = E_{\xi} F(x, \xi),$$

where  $\xi$  is a random vector with distribution  $P$  over a set  $\Xi$ . Also the special case:

$$f(x) = \frac{1}{m} \sum_{i=1}^m f_i(x),$$

where each  $f_i$  is convex and nonsmooth.

# Subgradients

For each  $x$  in domain of  $f$ ,  $g$  is a *subgradient of  $f$  at  $x$*  if

$$f(z) \geq f(x) + g^T(z - x), \quad \text{for all } z \in \text{dom } f.$$

Right-hand side is a *supporting hyperplane*.

The set of subgradients is called the *subdifferential*, denoted by  $\partial f(x)$ .

When  $f$  is differentiable at  $x$ , have  $\partial f(x) = \{\nabla f(x)\}$ .

We have strong convexity with modulus  $\mu > 0$  if

$$f(z) \geq f(x) + g^T(z - x) + \frac{1}{2}\mu\|z - x\|^2, \quad \text{for all } x, z \in \text{dom } f \text{ with } g \in \partial f(x).$$

# “Classical” Stochastic Approximation

Consider  $f(x) = \frac{1}{m} \sum_{i=1}^m f_i(x)$ .

At each iteration, choose  $\xi \in \{1, 2, \dots, m\}$  randomly, and use  $\nabla f_\xi(x)$  as the unbiased estimate of  $\nabla f$ .

Assume that there is  $M$  such that  $\|\nabla f_i(x)\|_2 \leq M$  for all  $x$  of interest.

**Basic SA Scheme:** At iteration  $k$ , choose  $\xi_k$  **i.i.d.** from  $\{1, 2, \dots, m\}$ , choose some  $\alpha_k > 0$ , and set

$$x^{k+1} = x^k - \alpha_k \nabla f_{\xi_k}(x^k).$$

When  $f$  is strongly convex, the analysis of convergence of  $E(\|x^k - x^*\|^2)$  is fairly elementary - see Nemirovski et al (2009). Steps  $\alpha_k = 1/(k\mu)$  lead to sublinear convergence

$$\frac{1}{2} E(\|x^k - x^*\|^2) \leq \frac{Q}{2k}, \quad \text{for } Q := \max \left( \|x^0 - x^*\|^2, \frac{M^2}{\mu^2} \right).$$



# What if $\mu$ is unknown, or zero?

The steplength choice  $\alpha_k = 1/(k\mu)$  requires knowledge of the modulus  $\mu$ . An underestimate of  $\mu$  can greatly degrade the performance of the method (see example in Nemirovski et al. 2009).

**Robust Stochastic Approximation** approach has a rate  $1/\sqrt{k}$  (in function value convergence), and works for weakly convex nonsmooth functions and is not sensitive to choice of parameters in the step length.

This is the approach that generalizes to *mirror descent* (Nemirovski et al, 2009).

At iteration  $k$ :

- set  $x^{k+1} = x^k - \alpha_k \nabla f_{\xi_k}(x^k)$  as before;
- define a **weighted average** of iterates so far:

$$\bar{x}^k = \frac{\sum_{i=1}^k \alpha_i x^i}{\sum_{i=1}^k \alpha_i}.$$

For any  $\theta > 0$  (not critical), choose step lengths to be

$$\alpha_k = \frac{\theta}{M\sqrt{k}}.$$

Then  $E[f(\bar{x}^k) - f(x^*)]$  converges to zero with rate approximately  $(\log k)/k^{1/2}$ . The choice of  $\theta$  is not critical.

# Robust Constant-Step Approach: HOGWILD!

(Nui et al., 2011)

- Set a target  $\epsilon$  for  $E[f(x^k) - f(x^*)]$ ;
- Estimate convexity modulus  $\mu$ , bound  $M$ , Lipschitz constant  $L$  for  $\nabla f$ ;
- Choose  $\vartheta \in (0, 1)$  and set

$$\alpha_k \equiv \frac{\vartheta \epsilon \mu}{2LnM^2}$$

Then have  $E[f(x^k) - f(x^*)] \leq \epsilon$  for all

$$k \geq \frac{2LnM^2 \log(L\|x^0 - x^*\|^2/\epsilon)}{\mu^2 \vartheta \epsilon}.$$

Unlike the basic SA scheme, this one is robust to knowledge of the convexity modulus  $\mu$ : an underestimate of  $\mu$  yields only a linear increase in the number of iterations.

Obtain a  $1/k$  rate *except for the log term*. We can remove this term by implementing a “backoff” scheme — periodically reduce the setpoint for  $\alpha_k$  by a factor  $\beta \in (0, 1)$ .

# Parallelizing HOGWILD!

Assume that each  $f_i$  depends on only a few components of  $x$  — say  $e_i \subset \{1, 2, \dots, n\}$ . Assume that  $e_i$  are generally small, and generally do not overlap too much.

Parallel implementation assumes that  $x$  is stored centrally, and updated by numerous processors that run the following loop (asynchronously):

- sample  $\xi_i \in \{1, 2, \dots, m\}$  uniformly;
- read the  $e_i$  components of  $x$  from central storage and evaluate  $\nabla f_i(x)$
- choose  $j \in e_i$  and update  $x_j \leftarrow x_j - \alpha[\nabla f_i(x)]_j$ ;

Assume that the “lag” in iterations between when any processor reads  $x$  and updates it is bounded by  $\tau$ .

The steplength strategy outlined above still works for the parallel variant, with similar convergence rate. The expressions for  $\alpha$  and number of steps  $k$  incorporate various quantities that characterize index sets  $\{e_1, e_2, \dots, e_m\}$ .

# HOGWILD! Computations

The parallel asynchronous version has been implemented on a dual Xeon machine (6 cores each  $\times$  2 hyperthreading). Used to solve very large sparse support vector machines problems on large data sets.

Matrix completion:

- Netflix:  $17,770 \times 480,189$  with 100M nonzeros;
- KDD Cup:  $625,000 \times 1M$  with 252M nonzeros;
- synthetic Jumbo problem:  $10M \times 10M$  with 2G nonzeros.

20 passes through the data, adjusting  $\alpha$  between epochs.

- About 2.5 hours for Jumbo problem.
- Up to 10 threads implemented; speedups up to 7 observed.

# Regularized Dual Averaging

(Nesterov, 2009) For  $\min f(x)$  with  $f$  convex, nonsmooth. Use subgradients  $g_i \in \partial f(x_i)$  and **average**, to obtain

$$\bar{g}_k = \frac{1}{k} \sum_{i=1}^k g_i.$$

Step:

$$x^{k+1} := \min_x \bar{g}_k^T x + \frac{\gamma}{\sqrt{k}} \|x - x^0\|_2^2, \quad \text{for some constant } \gamma > 0.$$

Possibly average the iterates  $x^1, x^2, x^3, \dots$  too.

Extended to  $\min \frac{1}{T} \sum_{t=1}^T f_t(x) + \tau c(x)$  by Xiao (2010).

# Manifold Identification in RDA

Under convexity assumptions, can show that  $\bar{g}_k$  approaches  $\nabla f(x^*)$  in expectation, with decreasing variance, at rate  $k^{-1/4}$ . (Faster if  $f$  is strongly convex).

Thus, under the usual assumptions of partial smoothness of  $\phi_\tau$  and nondegeneracy at  $x^*$ , a dense sequence of (non-averaged) iterates  $\{x^k\}$  eventually stays on the optimal manifold  $\mathcal{M}$ , with probability  $1 - O(k^{-1/4})$ .

(The  $O(\cdot)$  constant does not depend on problem dimension  $n$ .)

Motivates a 2-phase algorithm in which we switch to a different strategy (e.g. approximate reduced Newton method) on the low-dimensional manifold identified by RDA.

(Lee, Wright, 2011)

# Conclusions

- Have sketched only some of the computational techniques relevant to sparse optimization.
- A thriving and fully interdisciplinary field.
- The literature is expanding rapidly, with researchers from the “applications” areas making fundamental contributions to optimization theory and algorithms.




# Conclusions

- Have sketched only some of the computational techniques relevant to sparse optimization.
- A thriving and fully interdisciplinary field.
- The literature is expanding rapidly, with researchers from the “applications” areas making fundamental contributions to optimization theory and algorithms.

**THANKS!**

# A Very Incomplete Bibliography

- M. Afonso, J. Bioucas-Dias, and M. Figueiredo, “An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems,” *IEEE Transactions on Image Processing* 20 (2011), pp. 681–695.
- A. Beck and M. Teboulle, “A fast iterative shrinkage-threshold algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences* 2 (2009), pp. 183–202.
- J. Bioucas-Dias and M. Figueiredo, “Multiplicative noise removal using variable splitting and constrained optimization,” *IEEE Transactions on Image Processing*, 19 (2010), pp. 1720–1730.
- R. Byrd, G. Chin, W. Neveitt, and J. Nocedal, “On the use of stochastic Hessian information in unconstrained optimization,” 2010.
- J.-F. Cai, E. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” Technical Report, 2008.
- A. Chambolle, “An algorithm for total variation minimization and applications,” *Journal of Mathematical Imaging and Visualization*, 20 (2004), pp. 89–97.
- V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, “The convex geometry of linear inverse problems,” Technical Report, December 2010.
- S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing* 20 (1998), pp. 33–61.
- P. Combettes and V. Wajs, “Signal recovery by proximal forward-backward splitting,” *Multiscale Modeling and Simulation* 4 (2005), pp. 1168–1200.
- Ding, S., Wahba, G., and Zhu, X., “Learning higher-order graph structure with features by structure penalty,” in preparation, June 2011.
- J. Eckstein and D. Bertsekas, “On the Douglas-Rachford splitting method and the proximal-point algorithm for maximal monotone operators,” *Mathematical Programming* 

- M. Figueiredo, R. Nowak, and S. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE Journal on Selected Topics in Signal Processing*, 1 (2007), pp. 586–597.
- D. Goldfarb, S. Ma, and K. Scheinberg, “Fast alternating linearization methods for minimizing the sum of two convex functions,” submitted, 2010.
- T. Goldstein and S. Osher, “The split Bregman method for  $\ell_1$  regularized problems,” *SIAM Journal on Imaging Sciences* 2 (2009), pp. 323–343.
- E. Hale, W. Yin, and Y. Zhang, “A fixed-point continuation method for  $\ell_1$ -minimization: Methodology and convergence,” *SIAM Journal on Optimization* 19 (2008), pp. 1107–1130.
- M. Hintermüller and K. Kunisch, “Total bounded variation regularization as a bilaterally constrained optimization problem,” *SIAM Journal on Applied Mathematics* 64 (2004), pp. 1311–1333.
- M. Hintermüller and G. Stadler, “An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration,” *SIAM Journal on Scientific Computing* 28 (2006), pp. 1–23.
- R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, “Proximal methods for sparse hierarchical dictionary learning,” ICML, 2010.
- R. Jenatton et al., “Multi-scale mining of fMRI data with hierarchical structured sparsity,” Technical report HAL-INRIA-00589785, 2011.
- P. Lauzier, J. Tang, G.-H. Chen, “Prior image constrained compressed sensing (PICCS),” submitted, 2011.
- J. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J. Tropp, “Practical large-scale optimization for max-norm regularization,” NIPS, 2010.

- S. Lee and S. Wright, "Implementing algorithms for signal and image reconstruction on graphical processing units," 2008.
- S. Lee and S. Wright, "Manifold Identification of Dual Averaging Methods for Regularized Stochastic Online Learning," ICML, 2011.
- A. Lewis and S. Wright, "A proximal method for composite minimization," 2008.
- M. Lustig, "Sparse MRI," Ph.D. Thesis, Stanford University, 2008.
- J. Mairal, R. Jenatton, G. Obozinski, and F. Bach, "Network flow algorithms for structured sparsity," NIPS 2010.
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, 19 (2009), pp. 1574–1609.
- Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, 2004.
- Y. Nesterov, "Primal-dual subgradient methods for convex programs," *Mathematical Programming B* 120 (2009), pp. 221–259.
- F. Niu, B. Recht, C. Ré, and S. Wright, "HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent," Technical report, June, 2011.
- N. Rao, R. Nowak, S. Wright, N. Kingsbury, "Convex approaches to model avelet sparsity," ICIIP, 2011.
- B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions to linear matrix equations via nuclear norm minimization," *SIAM Review* 52 (2010), pp. 471–501.
- H. Robbins and S. Monro, "A Stochastic approximation method," *Annals of Mathematical Statistics* 22 (1951), pp. 400–407.
- L. Rudin, S. Osher, E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D* 60 (1992), pp. 259–268.
- W. Shi, G. Wahba, S. Wright, K. Lee, R. Klein, and B. Klein, "LASSO-Patternsearch algorithm with application to ophthalmology data," *Statistics and its Interface* 1 (2008).

- R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society B* 58 (1996), pp. 267–288.
- P. Tseng and S. Yun, "A coordinate descent method for nonsmooth separable minimization," *Math Prog B* (2009), pp. 387–423.
- B. Turlach, W. Venables, and S. Wright, "Simultaneous variable selection," *Technometrics* 47 (2005), pp. 349–363.
- E. van den Berg and M. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM Journal of Scientific Computing*, 31 (2008), pp. 890–912.
- S. Wright, "Accelerated Block-Coordinate Relaxation for Regularized Optimization," 2010.
- S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing* 57 (2009), pp. 2479–2493.
- L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *Journal of Machine Learning Research* 11 (2010), pp. 2543–2596.
- J. Yang and Y. Zhang, "Alternating direction algorithms for  $\ell_1$  problems in compressive sensing," CAAM Technical Report TR09-37, 2010.
- M. Zhu, S. J. Wright, and T. F. Chan, "Duality-based algorithms for total variation image restoration," *Computational Optimization and Applications*, 47 (2010), pp. 377–400.