Gradient Methods for Regularized Optimization

Stephen Wright

University of Wisconsin-Madison

ICIAM, Vancouver, July 2011

Outline

Sketch some computational optimization techniques for sparse / regularized optimization.

- Regularized formulations: ℓ_1 and other regularizers.
- Prox-Linear methods
  - Implementation for different regularizers
  - Extensions and enhancements
  - Identifying the optimal manifold.
- Stochastic gradient methods.
  - HOGWILD!
  - Regularized Dual Averaging

Includes joint work with S. Lee, A. Lewis, B. Recht, C. Ré, F. Niu.

Sparse / Regularized Optimization

Many applications need structured, approximate solutions of optimization problems, 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 $\Rightarrow$ need to gather / sample less data to capture it.

How do we formulate and solve problems to promote the desired structure in the solutions? Depends on the context and structure of the application, but there are important common threads e.g. nonsmooth regularizers.

Examples: compressed sensing, machine learning, many application areas in medical imaging, geophysics, power grids, control.

ℓ_1 and Sparsity

Often seek a sparse approximate minimizer of f: one with few nonzeros.

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

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

ℓ_1-constrained form (variable selection): $\min f(x)$ subject to $\|x\|_1 \leq T$.

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

Compressed Sensing

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 true solution x^* is sparse: $\|x\|_0 \ll n$.*

The additional knowledge of sparsity makes it possible to find the exact solution x^* even when the system $Ax = y$ is underdetermined.

- Donoho (2006), Donoho & Tanner: estimate number of observations m in terms of ambient space dimension and number of nonzeros.
- Candès, Tao, Romberg (2004, 2006): Restricted Isometry (RIP).
- Zhang (2008): ℓ_1/ℓ_2 on random subspaces.
- Candès and Recht (2011): elementary analysis for Gaussian A.
- Chandrasekaran et al. (2010): Gaussian observations with general regularizers.

Other Structures, Other Regularizers

Aim to design the regularizer to induce a desired structure, while keeping the optimization problem tractable.

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 ℓ_∞" or "sum of ℓ_2" regularizers:

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

where $[k]$ (for $k = 1, 2, \ldots, m$) 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

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

Overlapping, with tree struture: In a wavelet representation $U = Wx$ of a natural image, coefficients x can be arranged in a quadtree, exposing a hierarchical relationship.

Can impose this structure explicitly (Baraniuk et al., 2010), or induce via group regularizers (Jenatton et al., 2010), (Rao et al., 2011).

Total-Variation Regularization

Given intensity measures U_{ij} for $i, j = 1, 2, \ldots, N$ (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:

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

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

Matrix Estimation / Completion

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

$$\Omega \subset \{(i,j) \,|\, i = 1,2,\ldots,m, \;\; j = 1,2,\ldots,n\}.$$

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

Desirable structures:

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

Example formulation:

$$\min_X \; \frac{1}{2}\|\mathcal{A}X - b\|_2^2 + \tau\|X\|_*.$$

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

A Fundamental Method: Prox-Linear

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}:

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

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

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

c vacuous $\Rightarrow$ reduces to gradient descent, with a line search.

c is an indicator function for a closed convex set X $\Rightarrow$ reduces to gradient projection.

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 ∇f. e.g. In compressed sensing, requires multiplication by A and A^T — inexpensive for partial FFT, wavelets, etc.

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 proximality operator (shrink operator) associated with c:

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

This operator is easy and cheap to compute for simple regularizers: $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$: $O(n)$ or $O(n \log n)$ operations.

Also easy for non-overlapping group norms (because of separability).

Calulating the Shrink

- **Tree-structured groups:** Do a partial ordering of the groups, smaller to larger, then apply shrink in order (Jenatton et al., 2010)
- **General overlapping groups:** More complicated. Can solve using duality and block-coordinate relaxation, gradient projections, or min-cost network flow (Mairal et al., 2010).
- **Total Variation:** Solve a denoising problem:

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

(Chambolle, 2004) (Zhu, Chan, Wright, 2009).

- **Nuclear Norm:**

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

Can solve explicitly using a singular value decomposition of Y^k, followed by an adjustment (shrink) of the singular values.
SVT: Compute a partial SVD (largest singular values) using Lanczos. (Cai, Candès, Shen, 2008)

A Prox-Linear Method

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

At iteration k:

- Solve for current α_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 α_k as needed until sufficient decrease is obtained:

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

- Increase α_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).

Variants / Enhancements / Extensions

- Basic IST: $\alpha_k \equiv \bar{\alpha} < 1/L$. Guarantees descent in ϕ_τ at every iteration.
- Nonmonotone method using a Barzilai-Borwein choice of parameter α_k (another SpaRSA variant).
- Continuation in the regularization parameter τ. Solve a sequence of problems for different τ, from large to small, and warm start.
- Block Coordinate Relaxation: Calculate just a partial gradient (subvector of ∇f) at each iteration. Align the partial gradients with group boundaries. (Tseng and Yun, 2009), (Wright, 2010)
- 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.
- Composite Minimization: $h(c(x))$ where $c : \mathbb{R}^n \to \mathbb{R}^m$ is smooth and $h : \mathbb{R}^m \to \mathbb{R}$ is prox-regular. (Lewis & Wright, 2008)

Optimal Manifold and its Identification

Given the solution x^*, the optimal manifold $\mathcal{M}$ is a well behaved surface passing through x^* such that

- it can be parametrized by a small number of variables (relative to n);
- The regularizer c behaves smoothly along $\mathcal{M}$.

Example: For ℓ_1 regularization, $\mathcal{M}$ is the set of points near x^* with the same nonzero elements as x^*:

$$\mathcal{M} = \{x \in \mathbb{R}^n \mid x_i^* = 0 \Rightarrow x_i = 0\}.$$

- Can be parametrized by $\|x^*\|_0$ variables.
- Near x^*, $\|\cdot\|_1$ is smooth — linear in fact:

$$-\sum_{x_i^*<0} x_i + \sum_{x_i^*>0} x_i.$$

Identification of $\mathcal{M}$

Some algorithms can identify the optimal manifold $\mathcal{M}$ without knowing x^*. Thus, can switch to a different method for searching on $\mathcal{M}$, which typically has much lower dimension than n.

Requires a nondegeneracy condition: Replace

$$\begin{aligned} \text{criticality:} \quad & 0 \in \partial \phi_\tau(x^*) \\ \text{by } \textit{strict} \text{ criticality:} \quad & 0 \in \operatorname{ri} \partial \phi_\tau(x^*). \end{aligned}$$

Prox-linear methods identify $\mathcal{M}$ from points sufficiently close to x^*. So does a regularized dual averaging method that we describe below....

Stochastic Gradient Algorithms

Solves min $f(x)$ where

- 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 ∂f.

Some Definitions: For each x in domain of f, g is a *subgradient of f at x* if

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

f is *strongly convex 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

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

where each function f_i depends only a small subset of the components of x, denoted by $[i] \subset \{1, 2, \ldots, n\}$. Thus $\nabla f_i(x)$ is nonzero only in the components $[i]$.

At iteration k, choose $\xi_k \in \{1, 2, \ldots, m\}$ randomly, and use $\nabla f_{\xi_k}(x)$ as the unbiased estimate of ∇f.

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

Basic SA and it Properties

At iteration k, choose ξ_k i.i.d. from $\{1, 2, \ldots .m\}$, 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, (with modulus μ) the analysis of convergence of $E(\|x^k - x^*\|^2)$ is fairly elementary (Nemirovski et al, 2009).

Steps $\alpha_k = 1/(k\mu)$ lead to sublinear convergence

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

To reduce $E(\|x^k - x^*\|^2)$ to less than ϵ, need $O(1/\epsilon)$ iterations.

What if μ is unknown, or zero? Robust SA

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

Robust Stochastic Approximation works for weakly convex nonsmooth functions and is not sensitive to choice of parameters in the step length.

- set $x^{k+1} = x^k - \alpha_k \nabla f_{\xi_k}(x^k)$, with $\alpha_k = \frac{\theta}{M\sqrt{k}}$, for some $\theta > 0$ (not critical);
- define a weighted average of iterates so far:

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

Then $E[f(\bar{x}^k) - f(x^*)]$ converges to zero with rate approximately $(\log k)/k^{1/2}$.

Robust Constant-Step Approach: HOGWILD!

(Niu, Recht, Ré, Wright, 2011)

- Set a target ϵ for $E[f(x^k) - f(x^*)]$;
- Estimate convexity modulus μ, bound M, Lipschitz constant L for ∇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 μ: an underestimate of μ 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 α_k by a factor $\beta \in (0, 1)$.)

Parallelizing HOGWILD!

SGD seems inherently serial, but there are several parallel versions:

- Master-Worker (Bertsekas & Tsitsiklis, 1985)
- Round-Robin (Langford et al., 2009)
- Average between Runs (Zinkevich et al., 2010)

All require synchronization — parallelism degrades due to lock contention.

Idea: Get rid of locking! Allow different processors to update a centrally stored x vector independently of each other.

Each processor run this process (independently and unsynchronized):

1. Sample $\xi \in \{1, 2, \ldots, m\}$;
2. Read subvector $x_{[\xi]}$ and calculate $\nabla f_\xi(x)$;
3. For $v \in [\xi]$, update $x_v \leftarrow x_v - \alpha[\nabla f_\xi(x)]_v$;

Assume atomicity only of the single-component update in Step 3.
Updates can be old by the time they are made.

HOGWILD! Convergence

Performs best when the groups $[i]$ are small and don't overlap much. Define constants that quantify this:

$$\Omega := \max_{i=1,2,\ldots,m} |[i]|,$$

$$\Delta := \frac{1}{m} \max_{v=1,2,\ldots,n} |i = 1,2\ldots,m \,|\, v \in [i]|$$

$$\rho := \frac{1}{m} \max_{i=1,2,\ldots,m} \left|\hat{i} = 1,2,\ldots,m \,|\, [\hat{i}] \cap [i] \neq \emptyset\right|$$

Assume $\|\nabla f_i(x)\| \leq M$; $\mu I \preceq \nabla^2 f \preceq LI$. Assume that longest delay between reading x and updating it is τ steps.

$$\text{For } \; k \geq \frac{2LM^2(1 + 6\tau\rho + 6\tau^2\Omega\Delta^{1/2}) \log(L\|x^0 - x^*\|^2/\epsilon)}{\mu^2\epsilon},$$

then after k steps with constant stepsize α, we have $E\left[f(x^k) - f^*\right] \leq \epsilon$.

HOGWILD! Computations

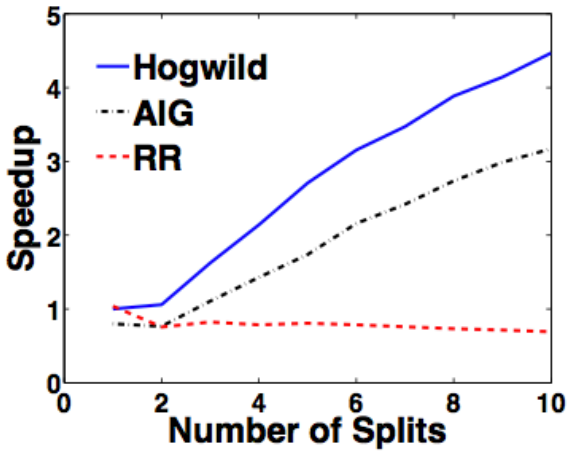
Shows speedups on a 10-core machine.

	data set	size (GB)	ρ	Δ	time (s)	speedup
SVM	RCV1	0.9	4.4E-01	1.0E+00	10	4.5
	Netflix	1.5	2.5E-03	2.3E-03	301	5.3
MC	KDD	3.9	3.0E-03	1.8E-03	878	5.2
	JUMBO	30	2.6E-07	1.4E-07	9,454	6.8
	DBLife	0.003	8.6E-03	4.3E-03	230	8.8
CUTS	Abdomen	18	9.2E-04	9.2E-04	1,181	4.1

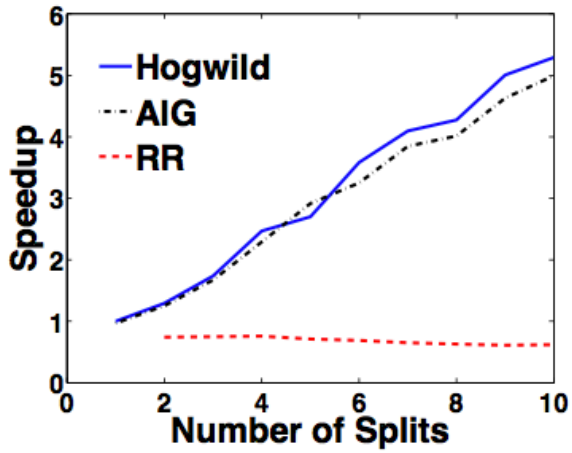
Round-Robin (Langford et al, 2009) is slower by a factor of 2-8.

HOGWILD! Speedups

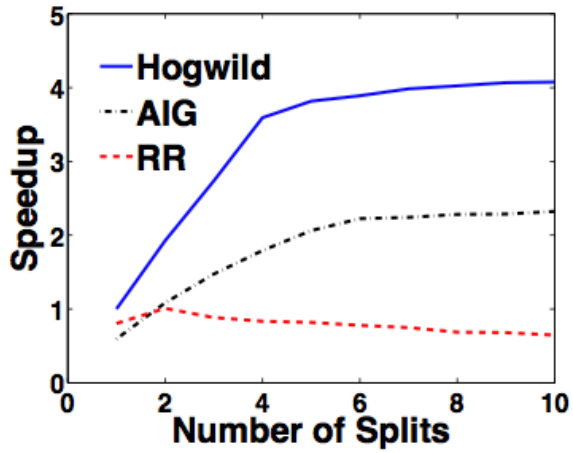
RR: Round-Robin. **AIG:** Same as HOGWILD! but locks the full subvector vector $[\xi]$ while updating. (HOGWILD! just locks individual indices.)



SVM
RCV1



MC
Netflix



CUTS
Abdomen

Regularized Dual Averaging

(Nesterov, 2009) For min $f(x)$ with f convex, nonsmooth.

Average all subgradients $g_i \in \partial f(x^i)$ visited so far, 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 primal iterates $x^1, x^2, x^3, \ldots$ too:

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

Extension to Regularized Formulation

Xiao (2010) extended the approach to the regularized, online setting

$$\min \phi_\tau(x) := \frac{1}{m} \sum_{t=1}^{m} f_t(x) + \tau c(x),$$

for which the subproblem is

$$x^{k+1} := \min_x \bar{g}_k^T x + \tau c(x) + \frac{\gamma}{\sqrt{k}} \|x - x^0\|_2^2.$$

Xiao proves convergence in expectation, for averaged iterates:

$$E\left[\phi_\tau(\bar{x}^k) - \phi_\tau(x^*)\right] \leq O(k^{-1/2}).$$

Manifold Identification

Assumes that x^* is a nondegenerate solution ($0 \in \mathrm{ri}\,[\nabla f(x^*) + \tau \partial c(x^*)]$) and a strong local minimizer *on the optimal manifold* $\mathcal{M}$.

Averaged gradients approach $\nabla f(x^*)$ in probability:

$$P\left(\|\bar{g}_k - \nabla f(x^*)\| > \epsilon\right) = O(\epsilon^{-2} k^{-1/4}).$$

Most of the sequence $\{x^k\}$ converges to x^*:

$$P\left(\|x^k - x^*\| > \epsilon\right) < O(\epsilon^{-2} k^{-1/4}) \text{ when } k \in \mathcal{S},$$

where for any k the sequence $\mathcal{S}$ contains all but a fraction of $O(k^{-1/4})$ of the elements of $\{1, 2, \ldots, k\}$.

For the same subsequence $\mathcal{S}$, we have for optimal manifold $\mathcal{M}$:

$$P(x^k \notin \mathcal{M}) \leq O(k^{-1/4}).$$

Constants in the $O(\cdot)$ terms don't depend on problem dimension.

Two-Phase Algorithm: RDA+

The manifold identification properties suggest a two-phase strategy:

- Use RDA to identify a reduced space containing optimal manifold $\mathcal{M}$;
- Run a different method on this reduced space, more suited to lower dimensions: more accurate gradients, reduced Newton-like steps.

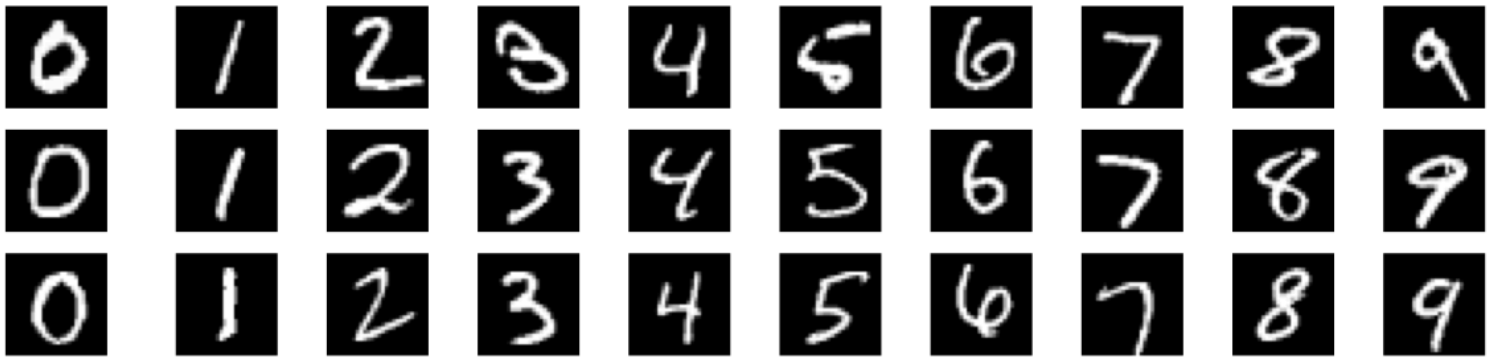
(Use a heuristic to decide when to switch.)

Implemented on ℓ_1 regularized logistic regression, where optimal manifold $\mathcal{M}$ is the set of points in $\mathbb{R}^n$ with the same nonzero structure as x^*.

Switch when the nonzeros of successive iterates x^k has settled down, and add zero components that are "close" to moving away from zero, to get a superset of $\mathcal{M}$. Run LPS (a prox-linear algorithm with reduced Newton steps) on the reduced space.

Tests with MNIST

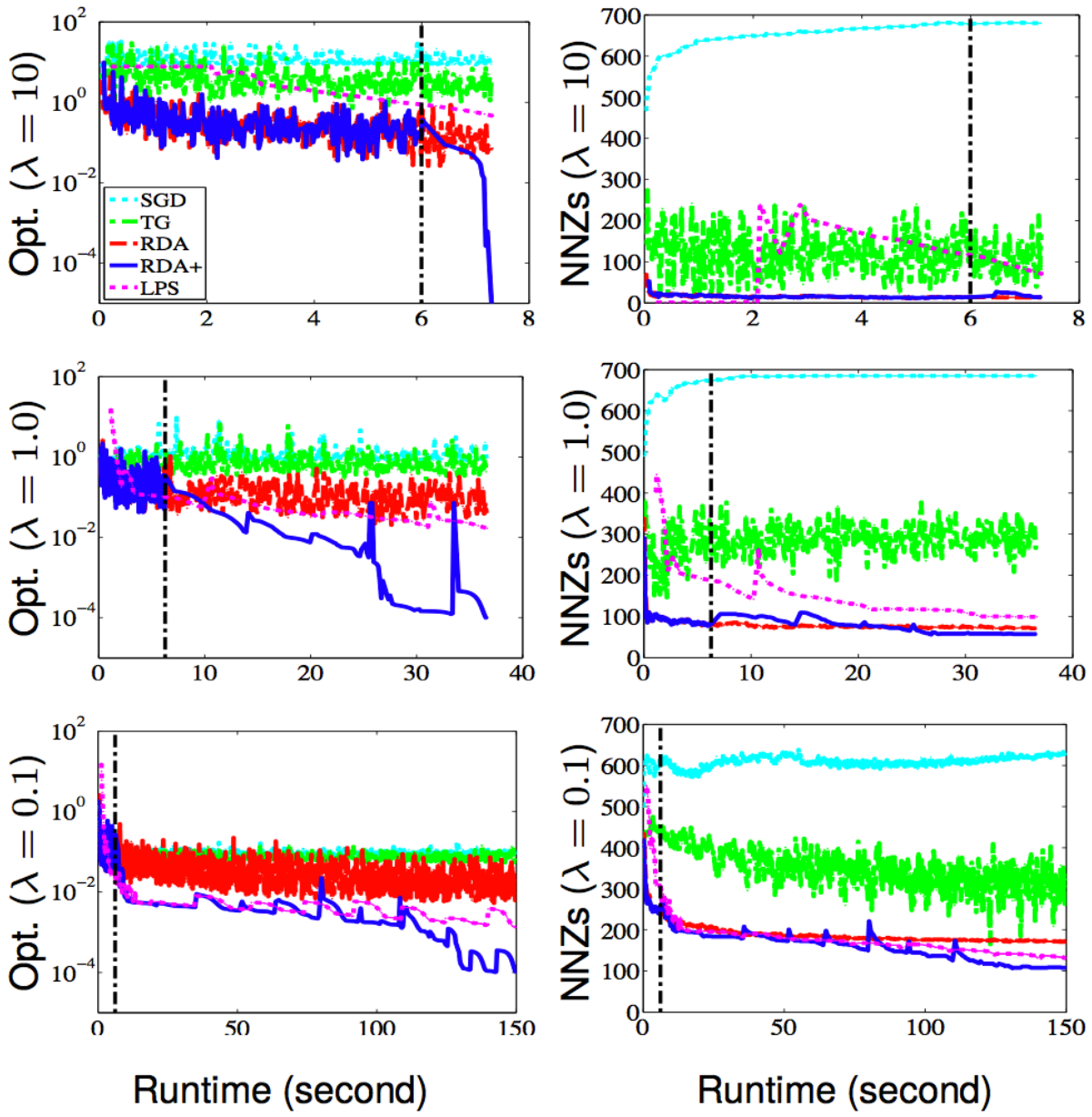
Standard data set in machine learning: identify handwritten digits.



Compare RDA+ with

- straight RDA
- straight LPS (on the full space)
- SGD
- Truncated Gradient ("shrinks" on every 10th step)

Comparing Runtimes and Sparsity



Stochastic Gradient References

- S. Lee and S. Wright, "Manifold Identification of Dual Averaging Methods for Regularized Stochastic Online Learning," ICML, 2011. Longer version submitted to JMLR, July 2011. (Posted to Optimization Online last night.)
- F. Niu, B. Recht, C. Ré, and S. Wright, "HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent," June, 2011. Submitted.

Summary

- Have discussed some recent work in regularized optimization and stochastic gradient methods.
- There's a confluence of interest between the two, particularly in machine learning.
- Many issues remain:
  - Design of regularizers to induce desired structure.
  - Enhancements of prox-linear that work better in practice.
  - Setting step length parameters in stochastic gradient.