Robust Regression and Efficient Optimization

Yaoliang Yu

University of Alberta

NICTA - Canberra
May 16, 2013

Table of Contents

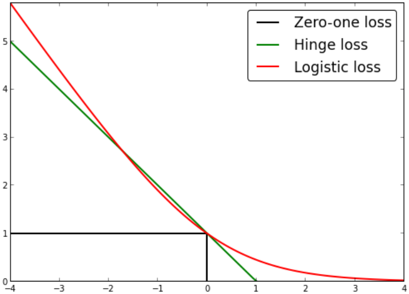
Coping with Hardness

Generic form for many ML problems:

$$\min_{w} f(w) + \lambda \cdot h(w).$$

Computationally challenging if

- the loss f is non-convex;
- the regularizer h is non-convex.



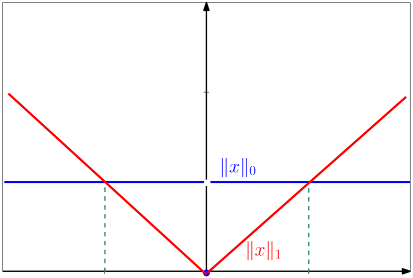


Table of Contents

Table of Contents

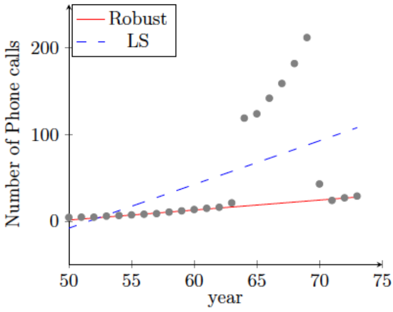
Introduction

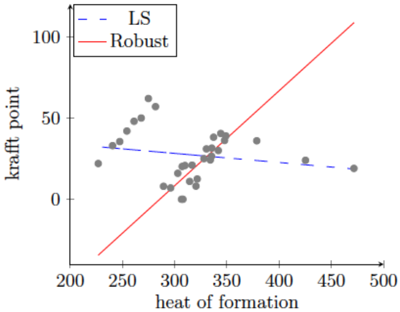
Problem:

- Real-world data is never clean;
- Even worse, often contains gross error.

Solutions:

- Two-stage: remove outliers first and then estimate parameters;
- One-stage: simultaneously achieve both.





Refs: (Rousseeuw-Leroy'87; Flores'11)

M-estimators and robust regression

Consider the linear regression model: $y = \langle \mathbf{x}, \mathbf{w} \rangle + \epsilon$.

- Given observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, want to estimate $\mathbf{w}$.

(Regularized) M-estimator: $\min_{\mathbf{w}} \sum_i \rho(y_i, \langle \mathbf{x}_i, \mathbf{w} \rangle) + \lambda \|\mathbf{w}\|_2^2$.

- Much is known if the loss ρ is convex.

Robustness: Would like the estimate to remain "reasonable" if perturb, say a single observation pair.

- Estimate remains bounded and away from boundary;
- Essentially requires nonzero breakdown point;
- Much is known if the loss ρ is bounded.

Refs: (Huber-Rochetti'09; Maronna-Martin-Yohai'06; etc)

State of the art

Properties	true or false				
M-estimator	1	1	1	0	1
Consistency	1	1	0	1	1
Robustness	1	0	1	1	1
Tractability	0	1	1	1	1
Achievable?	✓	✓	✓	?	✗

We proved that

1. If the loss ρ is convex, then ME cannot be robust;
2. If the loss ρ is bounded, then ME is NP-hard to find.

Isn't the ℓ_1 loss robust?

Argument:

- The median estimator is very robust;
- It minimizes the ℓ_1 loss: $\hat{m} \in \operatorname{argmin}_w \sum_{i=1}^{n} |w - y_i|$.

Caveat:

- $\mathbf{x}_i \equiv 1$ in the above example;
- Derivative of the obj: $\sum_i \rho'(y_i, \langle \mathbf{x}_i, w \rangle) \mathbf{x}_i - \lambda \mathbf{w}$.

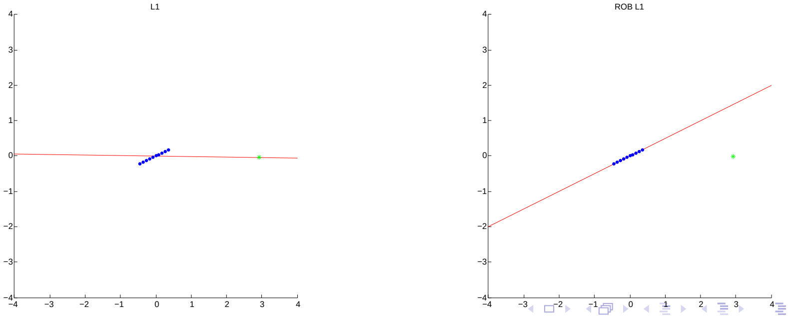
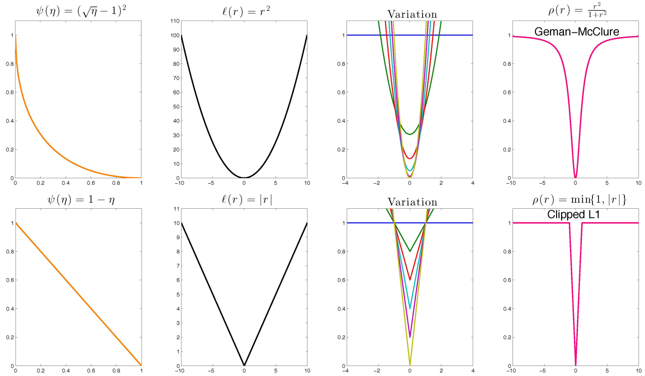


Table of Contents

Variational loss

$$\rho(x) = \min_{0 \le \eta \le 1} \eta \ell(x) + \psi(\eta).$$

- Includes most losses, even when ℓ and ψ are convex.



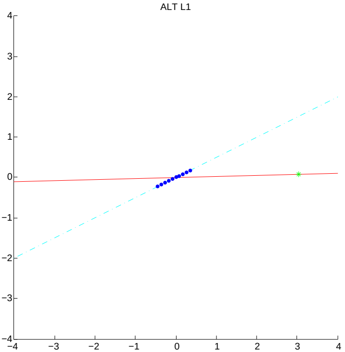
Refs: (Black-Rangarajan'96, Xu-Crammer-Schuurmans'06, etc.)

Variational M-estimator (Y-Aslan-Schuurmans'12)

Introduce outlier indicator $\boldsymbol{\eta}$:

$$\min_{\mathbf{w},\boldsymbol{\eta}\in[0,1]^n} \underbrace{\boldsymbol{\eta}^\top \boldsymbol{\ell}(\mathbf{y} - X\mathbf{w})}_{\text{loss on inliers}} + \underbrace{\mathbf{1}^\top \psi(\boldsymbol{\eta})}_{\text{penalize outliers}} + \underbrace{\tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1 \|\mathbf{w}\|_2^2}_{\text{regularizer}}$$

- If $(\mathbf{x}_i, y_i)$ incurs big loss, set $\eta_i = 0$ and suffer penalty $\psi(\eta_i)$;
- Otherwise set $\eta_i = 1$ and suffer no penalty;
- Not jointly convex in $\mathbf{w}$ and $\boldsymbol{\eta}$;
- Alternating can be bad.



Convex Relaxation

Reformulation:

$$\min_{\mathbf{w},\boldsymbol{\eta}\in[0,1]^n} \boldsymbol{\eta}^\top \ell(\mathbf{y} - X\mathbf{w}) + \mathbf{1}^\top \psi(\boldsymbol{\eta}) + \tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1\|\mathbf{w}\|_2^2$$

$$= \min_{\boldsymbol{\alpha},\boldsymbol{\eta}\in[0,1]^n} \boldsymbol{\eta}^\top \ell(\mathbf{y} - K\boldsymbol{\alpha}) + \mathbf{1}^\top \psi(\boldsymbol{\eta}) + \tfrac{\lambda}{2}\|\boldsymbol{\eta}\|_1 \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$$

$$= \min_{\boldsymbol{\eta}\in[0,1]^n} \max_{\boldsymbol{\nu}} \mathbf{1}^\top \psi(\boldsymbol{\eta}) - \boldsymbol{\eta}^\top(\ell^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^\top \left(K \circ (\boldsymbol{\eta}\|\boldsymbol{\eta}\|_1^{-1}\boldsymbol{\eta}^\top)\right) \boldsymbol{\nu}$$

$$= \min_{N\in\mathcal{N}_{\boldsymbol{\eta}}} \max_{\boldsymbol{\nu}} \mathbf{1}^\top \psi(\boldsymbol{\eta}) - \boldsymbol{\eta}^\top(\ell^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^\top (K \circ N) \boldsymbol{\nu},$$

Relaxation:

$$\mathcal{N}_{\boldsymbol{\eta}} = \{N : N \succeq 0, N\mathbf{1} = \boldsymbol{\eta}, \mathrm{rank}(N) = 1\}$$

$$\mathcal{M}_{\boldsymbol{\eta}} = \{M : M \succeq 0, M\mathbf{1} = \boldsymbol{\eta}, \mathrm{tr}(M) = 1\}$$

$$\geq \min_{M\in\mathcal{M}_{\boldsymbol{\eta}}} \max_{\boldsymbol{\nu}} \mathbf{1}^\top \psi(\boldsymbol{\eta}) - \boldsymbol{\eta}^\top(\ell^*(\boldsymbol{\nu}) - \Delta(\mathbf{y})\boldsymbol{\nu}) - \tfrac{1}{2\lambda}\boldsymbol{\nu}^\top (K \circ M) \boldsymbol{\nu}$$

Round: $\boldsymbol{\eta} = M\mathbf{1}$, re-solve $\mathbf{w}$.

Properties

Theorem (Tractability)

Convex-Concave program.

Theorem (Robustness)

Assume ℓ is Lipschitz and ψ' is bounded. Consider perturbation of the pair $(\mathbf{x}_1, y_1)$, the VM remains robust if either of the following holds

- y_1 *is bounded;*
- $\mathbf{x}_1$ *is bounded;*
- $\ell(y_1)/\|\mathbf{x}_1\|_2^2 \to \infty$.

Theorem (Consistency)

Assume ℓ is Lipschitz and ψ' is bounded. If the data consists of only inliers and outliers, then VM is (risk) consistent.

Some Experiment

- Seeded 5% outliers;
- RMSE (std) on *clean* test set.

Methods	Datasets							
	cal-housing		abalone		pumadyn		bank-8fh	
L2	1185	(124.59)	7.93	(0.67)	1.24	(0.42)	18.21	(6.57)
L1	1303	(244.85)	7.30	(0.40)	1.29	(0.42)	6.54	(3.09)
Huber	1221	(119.18)	7.73	(0.49)	1.24	(0.42)	7.37	(3.18)
LTS	533	(398.92)	755.1	(126)	0.32	(0.41)	10.96	(6.67)
GemMc	28	(88.45)	2.30	(0.01)	0.12	(0.12)	0.93	(0.80)
AltBndL1	1005	(603.00)	7.30	(0.40)	1.29	(0.42)	1.61	(2.51)
CvxBndL1	8	(0.28)	2.98	(0.08)	0.08	(0.07)	0.10	(0.07)
Gap(Cvx1)	0.005	(0.01)	0.001	(0.001)	0.267	(0.269)	0.011	(0.028)

Table of Contents

Conclusion

We have

- Showed the inherent dilemma between convexity and robustness;
- Developed the variational M-estimator.

Further questions:

- Approximation bound?
- Faster solver?

Table of Contents

Table of Contents

Conditional gradient (Frank-Wolfe'56)

Consider $$\min_{x \in C} \; f(x),$$

- C: compact convex;
- f: smooth convex.

1. $y_t \in \operatorname*{argmin}_{x \in C} \langle x, \nabla f(x_t) \rangle$;
2. $x_{t+1} = (1 - \eta) x_t + \eta y_t$.

(Frank-Wolfe'56; Canon-Cullum'68) proved that CG converges at $\Theta(1/t)$.

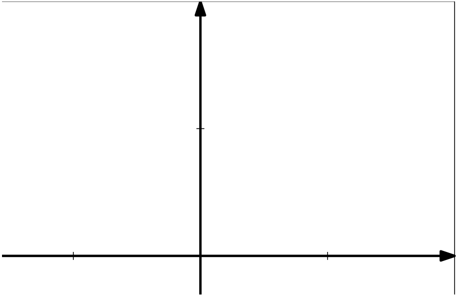
Gained much recent attention due to

- its simplicity;
- the greedy nature in step 1.

Refs: (Zhang'03; Clarkson'10; Hazan'08; Jaggi-Sulovsky'10; Bach'12; etc.)

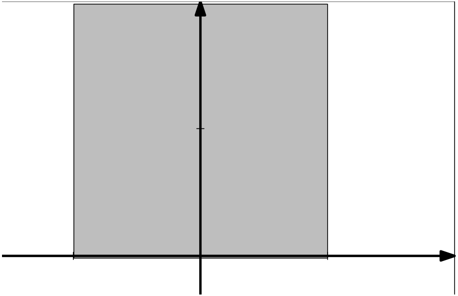
An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



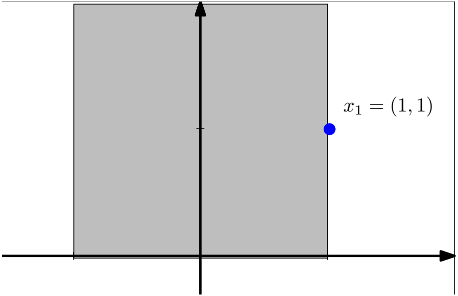
An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



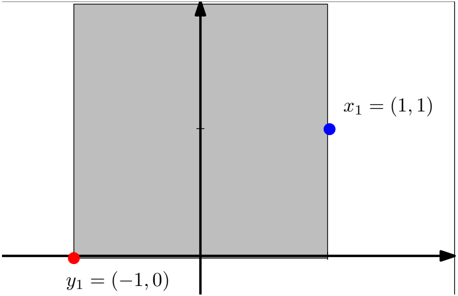
An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



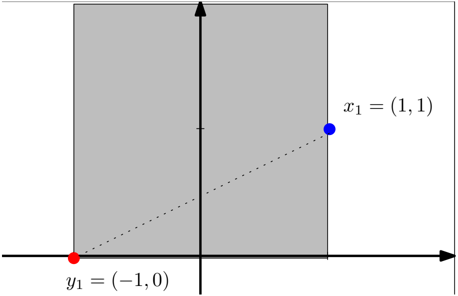
An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



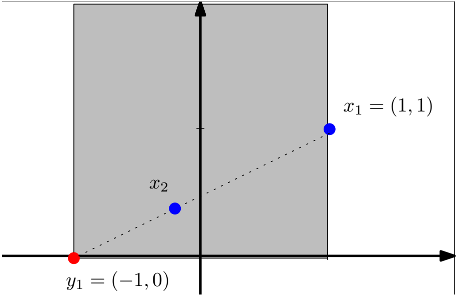
An Example

$$\min_{a,b} a^2 + (b+1)^2, \ \text{s.t.} \ |a| \le 1, 2 \ge b \ge 0$$



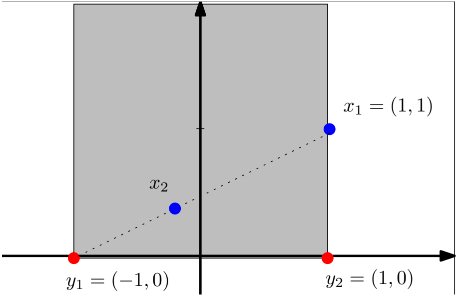
An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



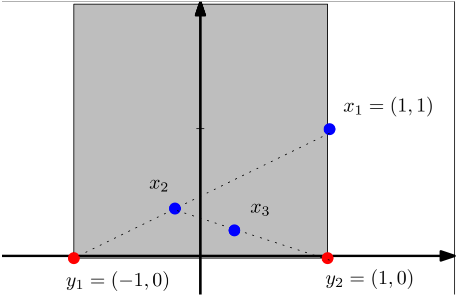
An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



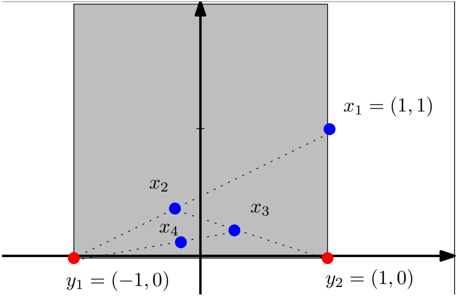
An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



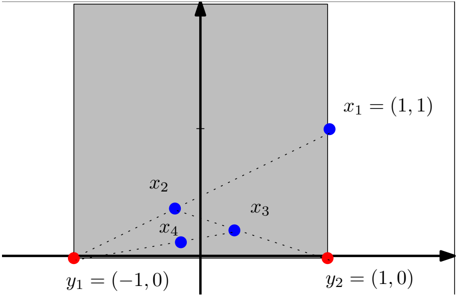
An Example

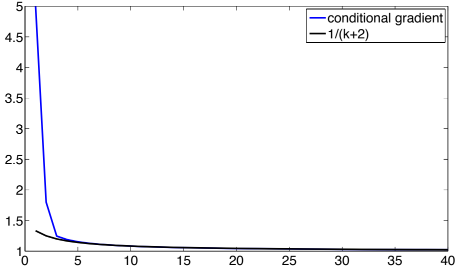
$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



An Example

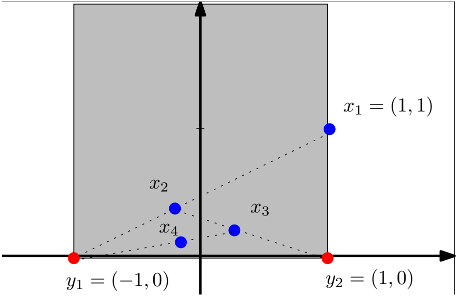
$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$

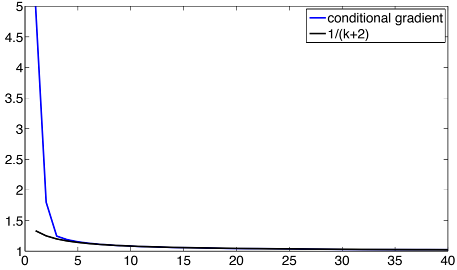




An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



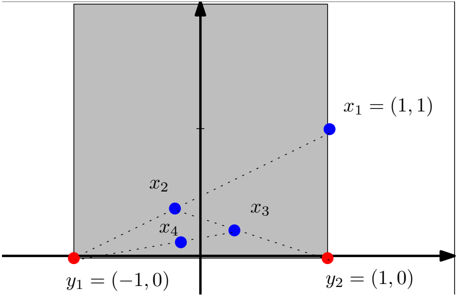


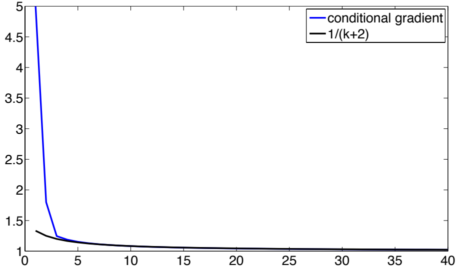
Can show $f(x_k) - f(x^\star) = 4/k + o(1/k)$.

Projected gradient converges in two iterations.

An Example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$





Can show $f(x_k) - f(x^\star) = 4/k + o(1/k)$.

Projected gradient converges in two iterations.

Refs: (Levtin-Polyak'66; Polyak'87; Beck-Teboulle'04) for faster rates.

The revival of CG: sparsity!

The revived popularity of conditional gradient is due to (Clarkson'10; Shalev-Shwartz-Srebro-Zhang'10), both focusing on

$$\min_{x:\ \|x\|_1 \le 1} f(x).$$

1. $y_t \leftarrow \operatorname*{argmin}_{\|y\|_1 \le 1} \langle y; \nabla f(x_t) \rangle,$ $\quad \texttt{card}(y_t) = 1;$
2. $x_{t+1} \leftarrow (1-\eta)x_t + \eta y_t,$ $\quad \texttt{card}(x_{t+1}) \le \texttt{card}(x_t) + 1.$

Explicit control of the sparsity. $\quad 1/\epsilon$ vs. $1/\sqrt{\epsilon}$.

Later on, (Hazan'08; Jaggi-Sulovsky'10) generalized the idea to SDPs.

Table of Contents

Generalized conditional gradient

Consider
$$\min_{x} \; f(x) + \lambda \cdot \kappa(x),$$

- f: smooth convex;
- κ: gauge (not necessarily smooth).

Important distinction:

- composite, with a non-smooth term;
- unconstrained, hence unbounded domain.

1. Polar operator: $y_t \in \operatorname*{argmin}_{x:\kappa(x)\leq 1} \langle x, \nabla f(x_t) \rangle$;
2. line search: $s_t \in \operatorname*{argmin}_{s \geq 0} f((1-\eta)x_t + \eta s y_t) + \lambda \eta s$;
3. $x_{t+1} = (1-\eta)x_t + \eta s_t y_t$.

Convergence Rate

$$\min_x \; f(x) + \lambda \cdot \kappa(x)$$

Theorem (Zhang-Y-Schuurmans'12)

If f and κ have bounded level sets and $f \in C^1$, then GCG converges at rate $O(1/t)$, where the constant is independent of λ.

Moreover, if using α-approximate PO, then GCG converges at rate $O(1/t)$ to an α-approximate solution.

- Proof is simple: Line search is as good as knowing $\kappa(x^*)$;
- Note that we upper bound $\kappa((1-\eta)x_t + \eta s y_t) \leq (1-\eta)\kappa(x_t) + \eta s$;
- Still too slow!

Local improvement

Assume some procedure (say BFGS) that can *locally* minimize the nonsmooth problem $\min_x f(x) + \lambda \cdot \kappa(x)$, or some variation of it.

Combine this local procedure with some globally convergent routine?

Two conditions:

- The local procedure cannot incur big overhead;
- Cannot ruin the globally convergent routine.

Both are met by the GCG.

Refs: (Burer-Monteiro'05; Mishra et al'11; Laue'12)

Case study: Matrix completion with trace norm

Consider
$$\min_X \frac{1}{2} \sum_{(i,j)\in\mathcal{O}} (X_{ij} - Z_{ij})^2 + \lambda \cdot \|X\|_{\mathrm{tr}}.$$

The only nontrivial step in GCG:

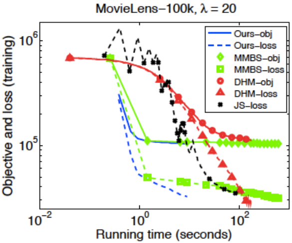
- Polar operator: $Y_t \in \operatorname{argmin}_{\|Y\|_{\mathrm{tr}} \le 1} \langle Y, G_t \rangle$, amounts to the dominating singular vectors of $-G_t$.

In contrast, popular gradient methods need the *full* SVD of $-G_t$.

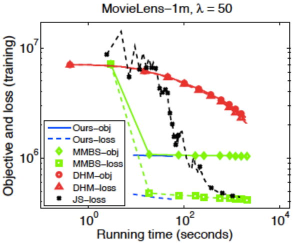
Variation (Srebro'05): $\frac{1}{2} \min_{U,V} \sum_{(i,j)\in\mathcal{O}} ((UV)_{ij} - Z_{ij})^2 + \lambda \cdot (\|U\|_F^2 + \|V\|_F^2)$.

- Not jointly convex in U and V;
- But smooth in U and V;
- Y_t in GCG is rank-1 hence $X_t = UV$ is of rank at most t.

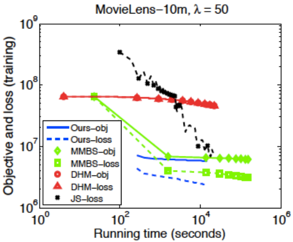
Case study: Experiment



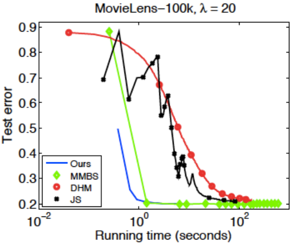
(a) Objective & loss vs time (loglog)



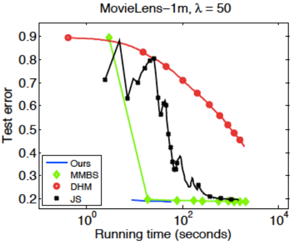
(a) Objective & loss vs time (loglog)



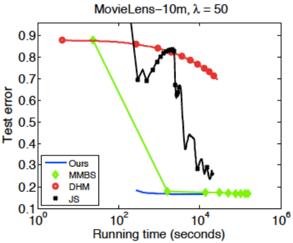
(a) Objective & loss vs time (loglog)



(b) Test NMAE vs time (semilogx)



(b) Test NMAE vs time (semilogx)



(b) Test NMAE vs time (semilogx)

Interpretation

Dictionary learning problem:

$$\min_{D \in \mathbb{R}^{m \times r}, \Phi \in \mathbb{R}^{r \times n}} L(X, D\Phi).$$

- Many applications: NMF, sparse coding ...
- Not *jointly* convex, in fact NP-hard for fixed r;

Convexify by *not* constraining the rank *explicitly*: relax r!

Refs: (Bengio et al'05; Bach-Mairal-Ponce'08; Zhang-Y-White-Huang-Sch'10)

Convexification

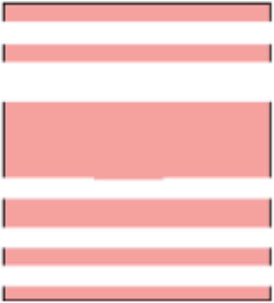
$$\min_{D,\Phi} L(X, D\Phi) + \lambda \cdot \Omega(\Phi).$$

- Let $D_{:i}$ have unit norm (say ℓ_2);
- Put row-wise norm on Φ: *implicitly* constraining the rank;
- Rewrite $\hat{X} := D\Phi = \sum_i \|\Phi_{i:}\| \cdot D_{:i} \frac{\Phi_{i:}}{\|\Phi_{i:}\|}$;
- Reformulate

$$\min_{\hat{X}} L(X, \hat{X}) + \lambda \cdot \kappa(\hat{X}) \quad \text{where}$$

$$\kappa(X) = \inf\{\textstyle\sum_i \sigma_i : X = \sum_i \sigma_i \cdot D_{:i} \frac{\Phi_{i:}}{\|\Phi_{i:}\|}\};$$

- Can apply GCG now, PO: $\min_{\mathbf{d},\phi} \mathbf{d}^\top G_t \frac{\phi}{\|\phi\|}$.



Setting both norms to ℓ_2, we recover the matrix completion example.

Table of Contents

Multiview (White-Y-Zhang-Schuurmans'12)

The complexity of GCG is packed into the PO:

$$\left\{ \min_{x:\kappa(x)\leq 1} \langle g, x \rangle \right\} = -\kappa^\circ(-g).$$

Recall that in the dictionary learning problem:

$$\left\{ \min_{\mathbf{d},\mathbf{w}} \mathbf{d}^\top G \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\} = -\left\{ \max_{\mathbf{d}} \|G^\top \mathbf{d}\|^\circ \right\}$$

In multiview learning, partition $\mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$ and constrain their norms resp..

Harder than single-view, but still doable:

$$\max_{\|\mathbf{d}_1\|=1, \|\mathbf{d}_2\|=1} \begin{bmatrix} \mathbf{d}_1^\top & \mathbf{d}_2^\top \end{bmatrix} GG^\top \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix} = \mathrm{tr}\left(GG^\top \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix} \begin{bmatrix} \mathbf{d}_1^\top & \mathbf{d}_2^\top \end{bmatrix} \right)$$

$$\frac{2(2+1)}{2} > 2.$$

Table of Contents

Conclusion

We have

- introduced the GCG;
- discussed efficient computations of PO;
- applied to MC, Group Lasso, etc.

Further questions

- nonsmooth?
- stochastic?

Thank you !