Fast Projections onto $\ell_{1,q}$-norm Balls
(for grouped feature selection)

SUVRIT SRA
Max Planck Institute for Intelligent Systems, Tübingen

ECML 2011, Athens, Greece.

Regularized Optimization

$$\boxed{\begin{array}{l} \ell(x) + \lambda r(x) \\ \ell(x) \quad \text{s.t. } r(x) \leq \gamma \end{array}}$$

Regularized Optimization

$$\boxed{\begin{array}{l} \ell(x) + \lambda r(x) \\ \ell(x) \quad \text{s.t. } r(x) \leq \gamma \end{array}}$$

Inverse problems	**Comp. Statistics**	Compr. Sensing	**Machine Learning**

Minimize

$$\boxed{\begin{array}{c} \ell(x)+\lambda r(x) \\ \text{or} \\ \ell(x) \quad \text{s.t. } r(x) \leq \gamma \end{array}}$$

Introduction - Background

Minimize

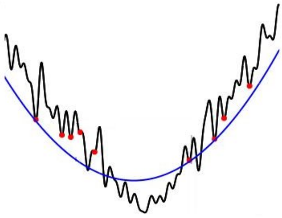
$$\ell(x) + \lambda r(x)$$

or

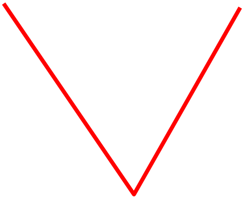
$$\ell(x) \quad \text{s.t. } r(x) \leq \gamma$$

We solve

$$\boxed{\ell(x) \quad \text{s.t.} \quad r(x) \leq \gamma}$$



Nonconvex, Differentiable



Non-differentiable

Main iteration

$$\boxed{x^{t+1} = \Pi(x^t - \alpha^t \nabla \ell(x^t))}$$

Main iteration

$$x^{t+1} = \Pi(x^t - \alpha^t \nabla \ell(x^t))$$

α^t : Barzilai-Borwein spectral step-size

Main iteration

$$\boxed{x^{t+1} = \Pi(x^t - \alpha^t \nabla \ell(x^t))}$$

α^t : Barzilai-Borwein spectral step-size

$$\boxed{\Pi(y) := \operatorname{argmin} \|x - y\|_2 \quad \text{s.t.} \quad r(x) \leq \gamma}$$

Step-size is closed form

Main work: **projection**

Mixed norms

Restrict to: $r(x) := \|x\|_{1,q}$

Parameter (sub)vectors $x_1, \ldots, x_T$

$$\|x\|_{1,q} := \sum_i \|x_i\|_q$$

Mixed norms

Restrict to: $r(x) := \|x\|_{1,q}$

Parameter (sub)vectors $x_1, \ldots, x_T$

$$\|x\|_{1,q} := \sum_i \|x_i\|_q$$

- Enforce joint sparsity / feature selection
- $q = 1$: ordinary ℓ_1; $q = 2$: group lasso;
- $q = \infty$: most severe—tends to eliminate entire feature
- multitask lasso; block compressed sensing; etc.

We need to solve

$$\boxed{\min \tfrac{1}{2}\|x - y\|_2^2 \quad \text{s.t.} \quad \|x\|_{1,q} \leq \gamma}$$

We need to solve

$$\boxed{\min \tfrac{1}{2}\|x - y\|_2^2 \quad \text{s.t.} \quad \|x\|_{1,q} \leq \gamma}$$

Lagrangian duality

There is a θ^* for which we can instead solve

$$\min \tfrac{1}{2}\|x - y\|_2^2 + \theta^*(\|x\|_{1,q} - \gamma)$$

This is easier!

Projections for mixed norms

$\|x\|_{1,q} = \sum_i \|x_i\|_q$; so problem *separable*

Separable ℓ_q-norm *proximity operations*

$$\min \tfrac{1}{2}\|x - y\|_2^2 + \theta^* \|x\|_q$$

$q = 1$: soft-thresholding (closed-form)
$q = 2$: soft-thresholding (closed-form)
$q = \infty$: requires reformulation

Finding θ^*

$$\boxed{\text{What about } \theta^*\text{?}}$$

Finding θ^*

What about θ^*?

Root finding

Let $g(\theta) := -\gamma + \|u(\theta)\|_{1,q}$

$u(\theta)$ via *proximity operator*

Finding θ^*

Theorem: $\theta^* \in [0, \theta_{\text{max}}]$

θ_{max} via *dual-norm*

Root finding: bisection, inv. quadratic, secant

Solution to ϵ-accuracy in $O(\log(\theta_{\text{max}}/\epsilon))$

$\ell_{1,\infty}$ projections

$$\boxed{\min \|X - Y\|_F^2 \quad \text{s.t. } \|X\|_{1,\infty} \leq \gamma}$$

Projection methods

Quattoni et al., ICML 2009 – **QP**

Our method, ECML 2011 – **FP**

$\ell_{1,\infty}$ projections

$c\gamma$	QP	FP
.50	1982s	31s

$\ell_{1,\infty}$ projections

$c\gamma$	QP	FP
.50	1982s	31s
.20	11491s	25s
.10	17064s	23s
.05	20165s	24s

Multitask Lasso

- Let X_j be data matrix for task j
- Seek parameter matrix W
- Columns corr. to tasks, rows to features
- Regularize *shared* features *across* tasks

$$\boxed{\sum_t \|y_t - X_t w_t\|_2^2 \ \text{ s.t. } \ \|W\|_{1,\infty} \leq \gamma}$$

Small to medium scale data

Size	# projs	SPG-QP	SPG-FP
27G	16	1054s	320s

Small to medium scale data

Size	# projs	SPG-QP	SPG-FP
27G	16	1054s	320s
0.8G	29	2474s	84s

Small to medium scale data

Size	# projs	SPG-QP	SPG-FP
27G	16	1054s	320s
0.8G	29	2474s	84s
8M	171	826s	31s

MTL on news data

- Subset of CMU News
- Five simultaneous feature selection tasks
- 54K features, 5 sets of 3K documents

MTL on news data

Density	SPG-QP	SPG-FP
.01	6800s	507s
.10	9759s	650s
.20	4746s	554s

SPG-QP spends 96% time in projections

SPG-FP spends 20% time in projections

Summary

- Efficient projections onto $\ell_{1,q}$-norm balls
- Application to mixed-norm regression
- Strong empirical results
- Several open issues remain

Summary

- Efficient projections onto $\ell_{1,q}$-norm balls
- Application to mixed-norm regression
- Strong empirical results
- Several open issues remain

Summary

- Efficient projections onto $\ell_{1,q}$-norm balls
- Application to mixed-norm regression
- Strong empirical results
- Several open issues remain