A smoothing approach for Composite Conditional Gradient with nonsmooth loss

Applications to collaborative filtering

Federico Pierucci

Joint work with
Zaid Harchaoui, Jérôme Malick

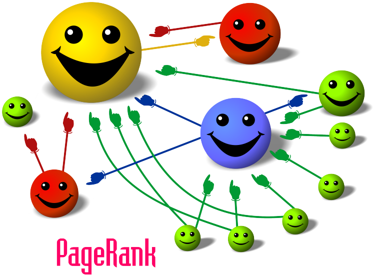
Laboratoire Jean Kuntzmann - Inria

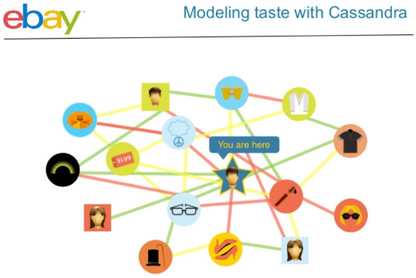
NYU - Courant Institute of Mathematical Sciences
New York, June 2015

Outline

Recommendation systems

- Related product recommendation (Amazon)
- Web page ranking (Google)
- Social recommendation (Facebook)
- Computational advertising (Yahoo!)
- $\rightarrow$ Movie recommendation (Netflix)





1

Collaborative Filtering for movie recommendation systems

Data: for user *i* and movie *j*
$X_{ij} \in \{0, 0.5, 1, 1.5, 2 \ldots, 4.5, 5\}$ ratings

The **aim** is to guess a future evaluation
$(i, j) \mapsto X_{ij} = ?$

Characteristics of collaborative filtering:

- **large scale**: size$(X) \sim 100,000 \, x \, 100,000$
- sparse data: size$(\mathcal{I}) << $ total entries of X
- no external data
- method is not content based

Outline

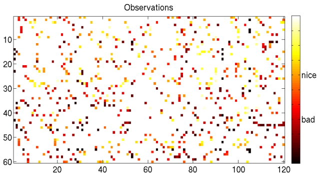
Convex optimization problem - Matrix completion

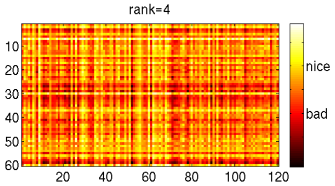
$$\min_{W \in \mathbb{R}^{d \times k}} \quad \frac{1}{N} \sum_{(i,j) \in \mathcal{I}} |W_{ij} - X_{ij}| + \lambda \|W\|_{\sigma,1}$$

where nuclear norm $\|W\|_{\sigma,1}$ is the sum of singular values of W
N = size($\mathcal{I}$) = Number of known entries of the matrix (=known rates)

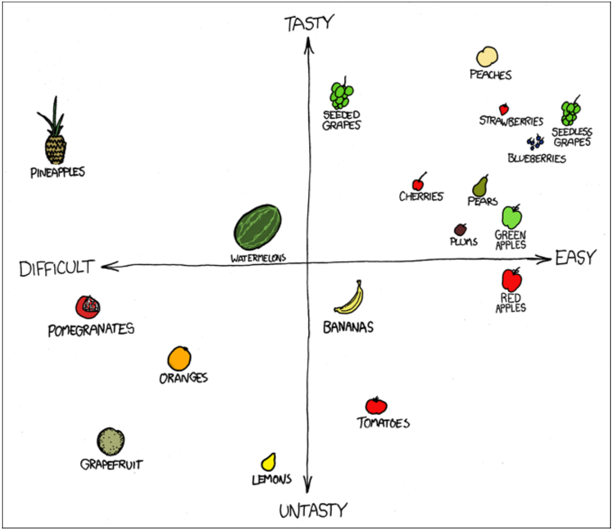
Why ℓ_1 loss? Previous work with $\|\cdot\|_2^2$ [Becker Bobin Candes 2009]
Here we consider ℓ_1 penality for more robustness to outliers.

Why nuclear-norm regularizer? Movies rates are supposed to be a linear combination of few "movie types" which are deduced observing only the ratings.





Why nuclear-norm regularizer?
Classes are embedded in a low dimension subspace of the feature space.



xkcd.com

Convex optimization problem

$$\underset{W \in \mathbb{R}^{d \times k}}{\text{minimize}} \quad R_{\text{emp}}(W) + \lambda \|W\|_{\sigma,1} \qquad \text{“doubly” nonsmooth problem}$$

- Algorithm: proximal algorithms (not scalable on large scale) [Nemirovski Yudin 1976] [Nesterov 2005]
- Issue: proximal operator related to nuclear-norm, requires computing the complete SVD of W.

Convex optimization problem

$$\underset{W \in \mathbb{R}^{d \times k}}{\text{minimize}} \quad R_{\text{emp}}(W) + \lambda \, \|W\|_{\sigma,1} \qquad \text{“doubly” nonsmooth problem}$$

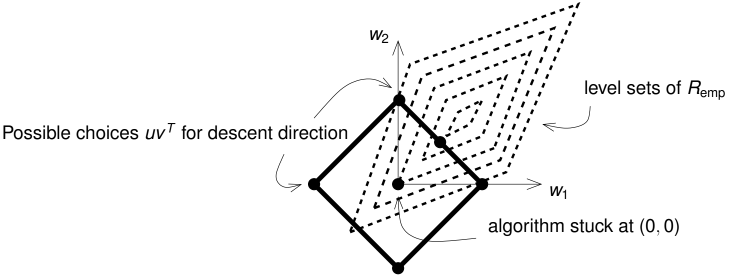
- Algorithm: proximal algorithms (not scalable on large scale) [Nemirovski Yudin 1976] [Nesterov 2005]
- Issue: proximal operator related to nuclear-norm, requires computing the complete SVD of W.

What if the loss were smooth?

$$\underset{W \in \mathbb{R}^{d \times k}}{\text{minimize}} \quad R_{\text{emp}}(W) + \lambda \, \|W\|_{\sigma,1} \qquad \text{with a smooth } R_{\text{emp}}$$

- Algorithm: Composite Conditional Gradient (scalable) [Harchaoui, Juditsky, Nemirovski, 2013]
  Requires to compute appropriate top singular vector pairs (an order of magnitude simpler than computing SVD)

Composite Conditional Gradient with nonsmooth loss does not converge



$$\text{minimize} \quad R_{\text{emp}}(W) + \lambda \|W\|_{\sigma,1}$$

Our approach:

- to smooth the loss (in a controllable way)
- to use Composite Conditional Gradient algorithms with smooth risk

Extension of Composite Conditional Gradient algorithms for doubly nonsmooth learning problems, e.g. collaborative filtering.

Outline

Smoothing of the loss function

Aim: build a family of smooth surrogates of R_{emp} parametrized by γ

$$\{R^{\gamma}_{\text{emp}}\}_{\gamma>0} \quad \text{with } R^{\gamma}_{\text{emp}} \text{ smooth}$$

Assumption:
The empirical risk is the support function of a convex compact set $\mathcal{B}$ in $\mathbb{R}^n$ (e.g. norms, gauge functions) composed with an affine function A

$$R_{\text{emp}}(W) = \max_{x\in\mathcal{B}} \langle x, AW\rangle$$

Construction of the family using the above structure: Fenchel-type γ-smooth function (adapted from [Nesterov 2005])

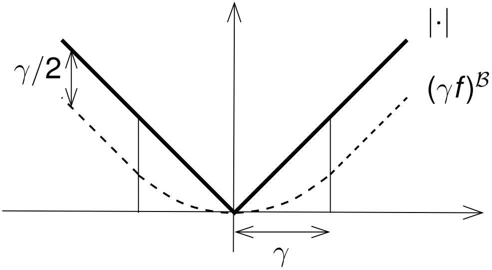
Definition (Fenchel-type γ-smooth function)

From Fenchel conjugate:

$$R^{\gamma}_{\text{emp}}(W) \quad := \quad \max_{x\in\mathcal{B}} \langle x, AW\rangle - \gamma f(x) \quad =: \quad (\gamma f)^{\mathcal{B}}(AW) \quad = \quad (\gamma f_{|\mathcal{B}})^*(AW)$$

$f : \mathcal{B} \to \mathbb{R}$ convex function

Example of smooth surrogate



$$|s| = \max_{x\in[-1,1]} xs$$

$$(\gamma f)^{\mathcal{B}}(s) = \max_{x\in[-1,1]} xs - \gamma\frac{1}{2}x^2$$

With $f(x) = \frac{1}{2}x^2$, we obtain the Huber function

$$(\gamma f)^{\mathcal{B}}(s) = \begin{cases} \frac{1}{2\gamma}s^2 & \text{if } |s| \leq \gamma \\ |s| - \frac{\gamma}{2} & \text{if } |s| > \gamma \end{cases} \qquad \nabla(\gamma f)^{\mathcal{B}}(s) = \begin{cases} 1 & \text{if } s > \gamma \\ \frac{1}{\gamma}s & \text{if } |s| \leq \gamma \\ -1 & \text{if } s < -\gamma \end{cases}$$

The parameter γ controls the approximation
Small $\gamma \Rightarrow$ better but less smooth approximation
Large $\gamma \Rightarrow$ worse but smoother approximation

Properties of the Fenchel-type γ-smooth function

Bounds of Fenchel-type γ-smooth function

- for all $x \in \mathcal{B} \quad m \leq f(x) \leq M \quad \Rightarrow$

$$\text{for all } s \in \mathbb{R}^k \quad \gamma m \leq \sigma(s) - (\gamma f)^{\mathcal{B}}(s) \leq \gamma M$$

- for $s \in \mathbb{R}^k \quad (\gamma f)^{\mathcal{B}}(s) \xrightarrow{\gamma \to 0} \sigma(s)$

The smooth surrogate can be made as tight as we want

Smoothness of $\mathcal{B}$-conjugate

f strongly convex on $\mathcal{B}$ (with constant 1)
then

- $(\gamma f)^{\mathcal{B}}$smooth
- $\nabla(\gamma f)^{\mathcal{B}}$ with Lipschitz constant $\frac{1}{\gamma}$ on $\mathbb{R}^k$

We now have the required smoothness to use Conditional Gradient algorithm

Outline

We use Composite Conditional Gradient algorithm fom [Harchaoui Juditsky Nemirovski 2013]

SCCG: Smoothed Composite Conditional Gradient

Inputs: λ, γ, ϵ
Initialize $W_0 = \mathbf{0}$
for $t = 1, \ldots, T(\epsilon)$ **do**
Call the oracle: $(u_t, v_t) = \underset{\|u\|_2=\|v\|_2=1}{\operatorname{argmin}} \langle \nabla R^{\gamma}_{\text{emp}}(W_{t-1}), uv^{\top} \rangle$

Compute

$$\min_{\theta_1,\ldots,\theta_t \geq 0} \quad R^{\gamma}_{\text{emp}} \left(\underbrace{\sum_{i=1}^{t} \theta_i u_i v_i^T}_{W_t} \right) + \lambda \sum_{i=1}^{t} \theta_i$$

Current solution $W_t = \sum_{i=1}^{t} \theta_i u_i v_i^T$
end for
Return W

⋆ **In theory**

Theorem (Complexity bound)

Set an optimization accuracy ϵ. Under some technical assumptions there is a smoothing parameter $\gamma(\epsilon) = O(\epsilon)$ such that after $T(\epsilon) = O(1/\gamma\epsilon)$ we have

$$R_{emp}(W_{T(\epsilon)}) - R^{\star}_{emp} \leq \epsilon$$

⋆ **In practice - choice of γ with grid search**

- Choose a family of smooth surrogate for loss (λ is fixed)
- Run SCCG for each γ on train set, fixed number of iteraitions
- Choose the best γ that minimizes R_{emp} on validation set

Outline

We recall

Original problem (doubly nonsmooth)

$$\underset{W\in\mathbb{R}^{d\times k}}{\text{minimize}} \quad \frac{1}{N}\sum_{(i,j)\in\mathcal{I}} |W_{ij}-X_{ij}| + \lambda \|W\|_{\sigma,1}$$

Surrogate problem with smooth loss

$$\underset{W\in\mathbb{R}^{d\times k}}{\text{minimize}} \quad \frac{1}{N}\sum_{(i,j)\in\mathcal{I}} \ell^{\gamma}(W_{ij}-X_{ij}) + \lambda \|W\|_{\sigma,1}$$

$\{W_t\}_t$ sequence of iterates from SCCG algorithm

Computations minimize the smoothed problem and return $\{W_t\}_t$

Data sets

MovieLens	users	movies	observations	sparsity
Small	943	1 682	100 000	6.3%
Medium	3 952	6 040	1 000 209	4.2%
Large	71 564	65 133	10 000 054	0.21%

Results Plot of the values of nonsmooth empirical risk $R_{emp}(W_t)$ for all three datasets [Pierucci, Harchaoui, Malick, Conférence d'Apprentissage Automatique Cap'2014]

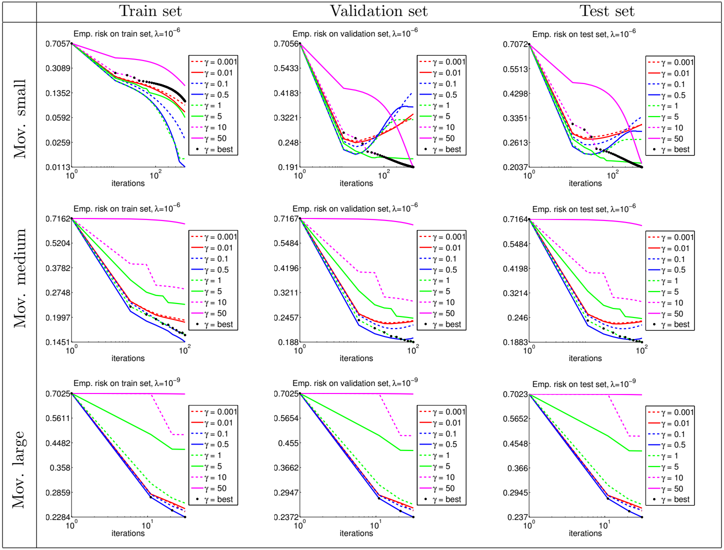


Figure 2: Movielens data - Empirical risk versus iterations.

Train set : we obtain sequences $\{W_t\}_t$ for a set of $\gamma \in [0.001, \ldots, 50]$ and $\lambda \in [0, \ldots, 10^{-2}]$

Validation set: we chose the parameters λ_{best} and γ_{best} which minimize $R_{emp}(W_t)$, at the last iteration

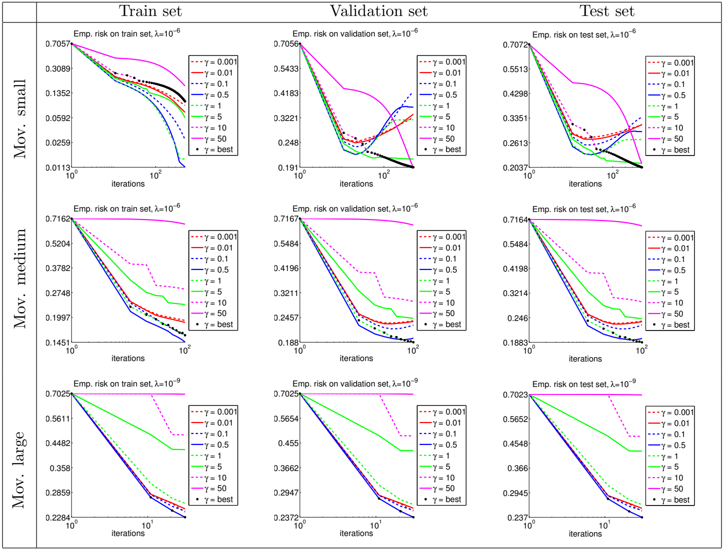


Figure 2: Movielens data - Empirical risk versus iterations.

Conclusion

- Collaborative Filtering with ℓ_1 loss
- Generalizable doubly nonsmooth objective function: nonsmooth loss + norm regularizer
- Algorithm SCCG suitable for large scale
- Efficient calibration of γ
- (To release) Matlab and python code - collaborative filtering for recommendation systems

Thank you for your attention

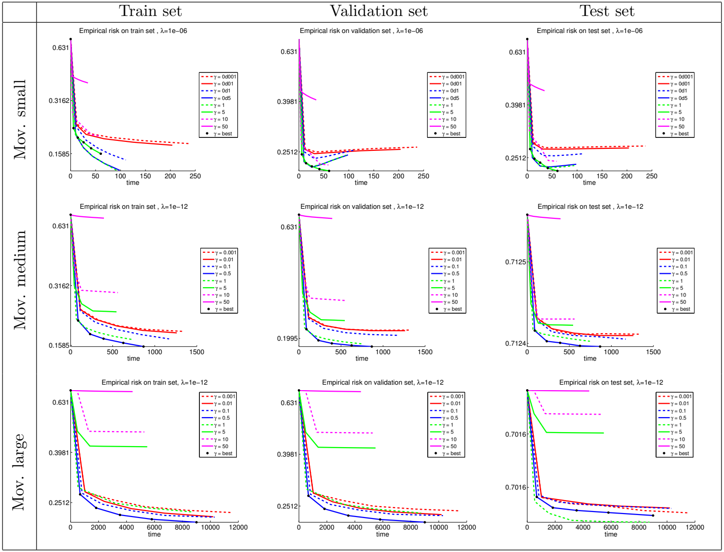


Figure 3: Movielens data - Empirical risk versus time. Related to all γ for the best choice of λ.

Key point: Consider the variable as weighted sum of atoms $a_i \in \mathcal{A}$

$$W = \sum_{i \in \mathcal{I}} \theta_i a_i, \quad \theta_i \in \mathbb{R}$$

SCCG - General version

Inputs: λ, γ, ϵ
Initialize $W_0 = \mathbf{0}$
for $t = 1, \ldots, T(\epsilon)$ **do**
Call the linear minimization oracle: $a_i = \mathbf{LMO}^{\gamma}(W_t)$
Compute
$$\min_{\theta_1,\ldots,\theta_t \geq 0} \lambda \sum_{i=1}^{t} \theta_i + R_{\text{emp}}^{\gamma}\left(\sum_{i=1}^{t} \theta_i a_i\right)$$
Current solution $W_t = \sum_{i=1}^{t} \theta_i a_i$
end for
Return $W = \sum_i \theta_i a_i$

Linear minimization operator (replaces the proximal operator)

$$\mathbf{LMO}^{\gamma}(W) := \operatorname*{argmin}_{a \in \mathcal{A}} \langle a, \nabla R_{\text{emp}}^{\gamma}(W) \rangle \ .$$