Αλγοριθμικές Μέθοδοι Βελτιστοποίησης με Έμφαση σε Κατανεμημένα Προβλήματα

Αλγόριθμοι Εγγύτατης Κλίσης

(Proximal Gradient Algorithms)

Δημήτρης Αμπελιώτης

Επίκουρος Καθηγητής, Ιόνιο Πανεπιστήμιο

Περιεχόμενα

- Το πρόβλημα της κυρτής ικανοποιησιμότητας (convex feasibility problem)
- Βελτιστοποίηση κυρτών συναρτήσεων με κυρτούς περιορισμούς – ο αλγόριθμος projected gradient
- Βελτιστοποίηση κυρτών μη-ομαλών συναρτήσεων
  - Γενίκευση της προβολής – τελεστής εγγύτητας
  - Ο αλγόριθμος εγγύτατης κλίσης
  - Ο αλγόριθμος εγγύτατης κλίσης με επιτάχυνση
- Συνήθεις όροι κανονικοποίησης και οι αντίστοιχοι τελεστές εγγύτητας
- Εφαρμογές
  - Αραιή αναπαράσταση σημάτων (ISTA, FISTA)
  - Συμπλήρωση πίνακα χαμηλής τάξης (SVT)
  - Διαχωρισμός υποβάθρου από προσκήνιο σε βίντεο (Robust PCA)
  - Μείωση θορύβου με διατήρηση ακμών (TV-Denoising)

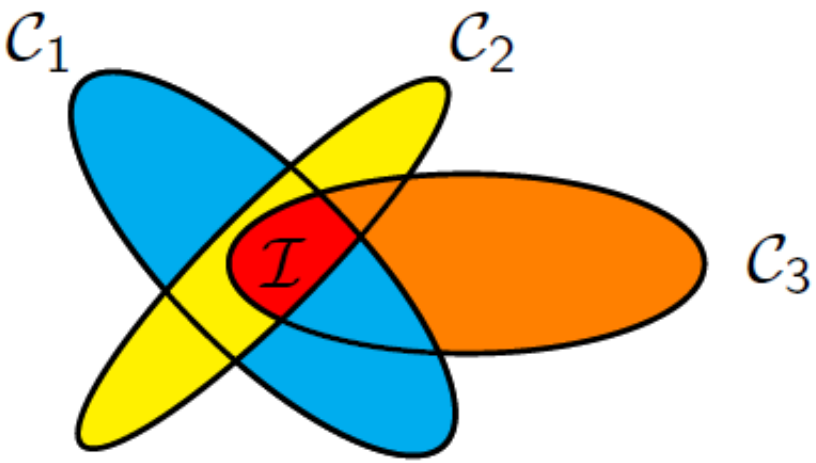
Το Πρόβλημα της Κυρτής Ικανοποιησιμότητας

(Convex Feasibility Problem)

Convex Feasibility Problem

- Το πρόβλημα της κυρτής ικανοποιησιμότητας συνίσταται στην εύρεση (υπολογισμό) ενός σημείου $\boldsymbol{p}$ το οποίο βρίσκεται στην τομή ενός συνόλου από κυρτά σύνολα

$$\boldsymbol{p} \in \mathcal{I} = \bigcap_{n=1}^{N} \mathcal{C}_n$$



Το πρόβλημα αυτό το συναντάμε σε προβλήματα ικανοποίησης πολλαπλών περιορισμών / συγχώνευσης γνώσης κ.λπ.

Convex Feasibility Problem

- Ορίζουμε τον τελεστή προβολής ενός σημείου $\boldsymbol{x} \in \mathbb{R}^d$ στο σύνολο $\mathcal{C}$ ως το σημείο εκείνο που ανήκει στο $\mathcal{C}$ και έχει ελάχιστη απόσταση από το σημείο $\boldsymbol{x}$, δηλαδή

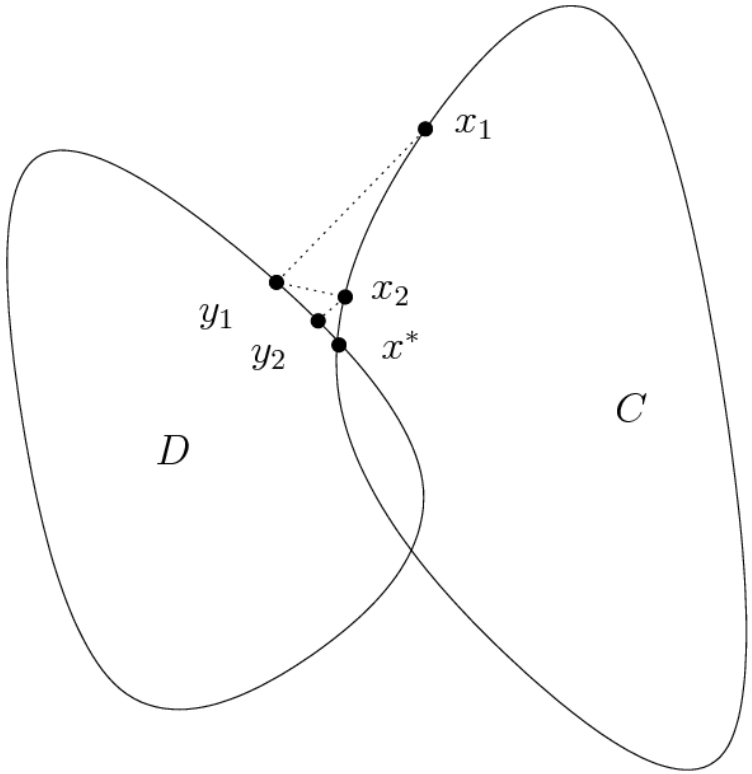
$$\Pi_{\mathcal{C}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}\in\mathcal{C}}(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$$

- Ένας ευρέως γνωστός αλγόριθμος που υπολογίζει ένα σημείο στην τομή N κλειστών, συμπαγών και κυρτών συνόλων, υποθέτοντας πως η τομή αυτή δεν είναι κενή, είναι ο αλγόριθμος Projection Onto Convex Sets (POCS), ο οποίος προβάλει διαδοχικά πάνω στα σύνολα $\mathcal{C}_n$

$\boldsymbol{x}^0 \in \mathbb{R}^d$
FOR $k = 0$ TO $+\infty$
$\quad i = k \bmod N$
$\quad \boldsymbol{x}^{k+1} = \Pi_{\mathcal{C}_{i+1}}(\boldsymbol{x}^k)$
END

Convex Feasibility Problem

- POCS: Παράδειγμα με δύο σύνολα



Convex Feasibility Problem

- Για ένα (κυρτό) σύνολο $\mathcal{C} \subset \mathbb{R}^d$ ορίζουμε μια *δείκτρια συνάρτηση (indicator function)* την οποία ορίζουμε ως

$$\delta_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} 0, & \alpha\nu\ \boldsymbol{x} \in \mathcal{C} \\ +\infty, & \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\alpha \end{cases}$$

- Έτσι, μια εναλλακτική μορφή για το πρόβλημα της κυρτής ικανοποιησιμότητας είναι η ακόλουθη

$$\boldsymbol{p} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left(\delta_{\mathcal{C}_1}(\boldsymbol{x}) + \delta_{\mathcal{C}_2}(\boldsymbol{x}) + \cdots + \delta_{\mathcal{C}_N}(\boldsymbol{x})\right)$$

Παρατηρούμε εύκολα πως τα σημεία που ανήκουν στην τομή των κυρτών συνόλων είναι τα μόνα που δίνουν τιμή 0 στη συνάρτηση κόστους, ενώ όλα τα σημεία εκτός της τομής δίνουν τιμή άπειρο

Ελαχιστοποίηση Κυρτής Συνάρτησης με Κυρτό Περιορισμό

(Αλγόριθμος Projected Gradient)

Convex Constrained Optimization

- Ας θεωρήσουμε το πρόβλημα βελτιστοποίησης

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{C}} \big(f(x)\big)$$

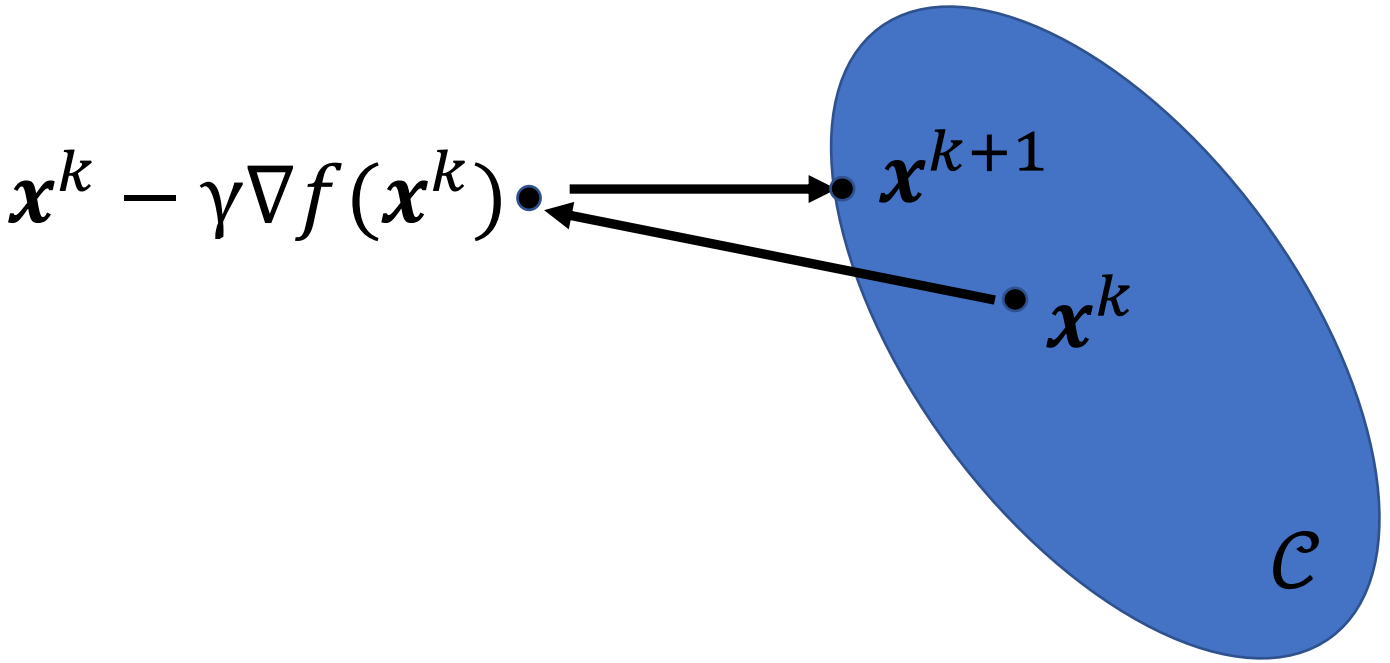
όπου $\mathcal{C} \subset \mathbb{R}^d$ ένα κυρτό, συμπαγές και κλειστό σύνολο και η συνάρτηση $f(x)$ είναι μια κυρτή συνάρτηση, παραγωγίσιμη (τουλάχιστον) μια φορά σε όλα τα σημεία του συνόλου $\mathcal{C}$

- Για το πρόβλημα αυτό, ένας ευρέως γνωστός αλγόριθμος βελτιστοποίησης είναι ο λεγόμενος Projected Gradient Αλγόριθμος

$\boldsymbol{x}^0 \in \mathcal{C}$
FOR $k = 0$ TO $+\infty$
$$\boldsymbol{x}^{k+1} = \Pi_{\mathcal{C}}\left(\boldsymbol{x}^k - \gamma\nabla f\left(\boldsymbol{x}^k\right)\right)$$
END

Convex Constrained Optimization

- Παράδειγμα
  - Αρχικά κινούμαστε προς την κατεύθυνση μείωσης της συνάρτησης κόστους
  - Αν βγούμε από το σύνολο $\mathcal{C}$, επιστρέφουμε στο πλησιέστερο σημείο εντός του συνόλου αυτού



Convex Constrained Optimization

- Χρησιμοποιώντας τον ορισμό για τη δείκτρια συνάρτηση ενός συνόλου, που είδαμε στα προηγούμενα, είναι δυνατόν να εκφράσουμε το πρόβλημα που εξετάζουμε εδώ σε μια εναλλακτική μορφή ως εξής

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \big(f(x) + \delta_{\mathcal{C}}(\boldsymbol{x})\big)$$

Παρατηρούμε πως η εισαγωγή της δείκτριας συνάρτησης $\delta_{\mathcal{C}}(\boldsymbol{x})$ στη συνάρτηση κόστους την κάνει να έχει τιμή άπειρο σε όλα τα σημεία που δεν ανήκουν στο σύνολο περιορισμού $\mathcal{C}$

Βελτιστοποίηση Σύνθετης Συνάρτησης

(με παραγωγίσιμη και μη-παραγωγίσιμη συνιστώσα)

Composite Convex Optimization

- Ενδιαφερόμαστε τώρα να ελαχιστοποιήσουμε συναρτήσεις κόστους που έχουν τη μορφή

$$F(x) = f(x) + g(x)$$

όπου η συνάρτηση $f(x)$ είναι κυρτή και παραγωγίσιμη, ενώ η συνάρτηση $g(x)$ είναι κυρτή αλλά μη-παραγωγίσιμη

- Στα προηγούμενα, είδαμε πως όταν εμφανιζόταν μια δείκτρια συνάρτηση $\delta_{\mathcal{C}}(\boldsymbol{x})$ ως όρος της συνάρτησης κόστους, η οποία δεν μπορεί να παραγωγιστεί, ο τελεστής προβολής χρησιμοποιούταν για να «ελαχιστοποιήσουμε» τον όρο αυτό
- Για συναρτήσεις $g(\boldsymbol{x})$ που δεν είναι παραγωγίσιμες, διαφορετικές γενικά από τις δείκτριες συναρτήσεις που ορίσαμε, μπορούμε να ορίσουμε μια γενίκευση του τελεστή προβολής, τον οποίο θα ονομάσουμε **τελεστή εγγύτατου σημείου (proximal operator)**
- Θα δούμε αλγορίθμους βελτιστοποίησης, που γενικεύουν αυτούς που είδαμε στα προηγούμενα (αυτοί που είδαμε αποτελούν εξειδικεύσεις)

Composite Convex Optimization

Γενίκευση του τελεστή προβολής – proximal operator

- Ο τελεστής προβολής μπορεί να οριστεί ως η λύση ενός προβλήματος βελτιστοποίησης της μορφής

$$\Pi_{\mathcal{C}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}\in\mathbb{R}^d} \left(\delta_{\mathcal{C}}(\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \right)$$

- Κατ' αναλογία, για μια (κυρτή) συνάρτηση $g(\boldsymbol{x})$ ορίζουμε τον τελεστή εγγύτατου σημείου (Proximal Operator) μέσω ενός προβλήματος βελτιστοποίησης της μορφής

$$prox_g(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}\in\mathbb{R}^d} \left(g(\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \right)$$

Όπως η προβολή ενός δοσμένου σημείου σε ένα συμπαγές κυρτό σύνολο είναι μοναδική, έτσι και το εγγύτατο σημείο ενός δοσμένου σημείου ως προς μια (όχι απαραίτητα ισχυρά) κυρτή συνάρτηση είναι μοναδικό

Composite Convex Optimization

- Ιδιότητες του Proximal Operator (χρησιμοποιούνται στις αποδείξεις σύγκλισης)
  - **Property 1:** Inclusion

$$\boldsymbol{p} = prox_g(\boldsymbol{x}) \Leftrightarrow \boldsymbol{x} - \boldsymbol{p} \in \vartheta g(\boldsymbol{p})$$

  - **Property 2:** Non-expansiveness

$$\|prox_g(\boldsymbol{x}) - prox_g(\boldsymbol{y})\|^2 + \left\|\left(\boldsymbol{x} - prox_g(\boldsymbol{x})\right) - \left(\boldsymbol{y} - prox_g(\boldsymbol{y})\right)\right\|^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

  - **Property 3:** Το σύνολο σταθερών σημείων (fixed points set) του τελεστή ταυτίζεται με το σύνολο ελαχιστοποιητών (minimizers) της συνάρτησης $g(\boldsymbol{x})$

Composite Convex Optimization

- Κατ' αναλογία με τον αλγόριθμο Projected Gradient Descend

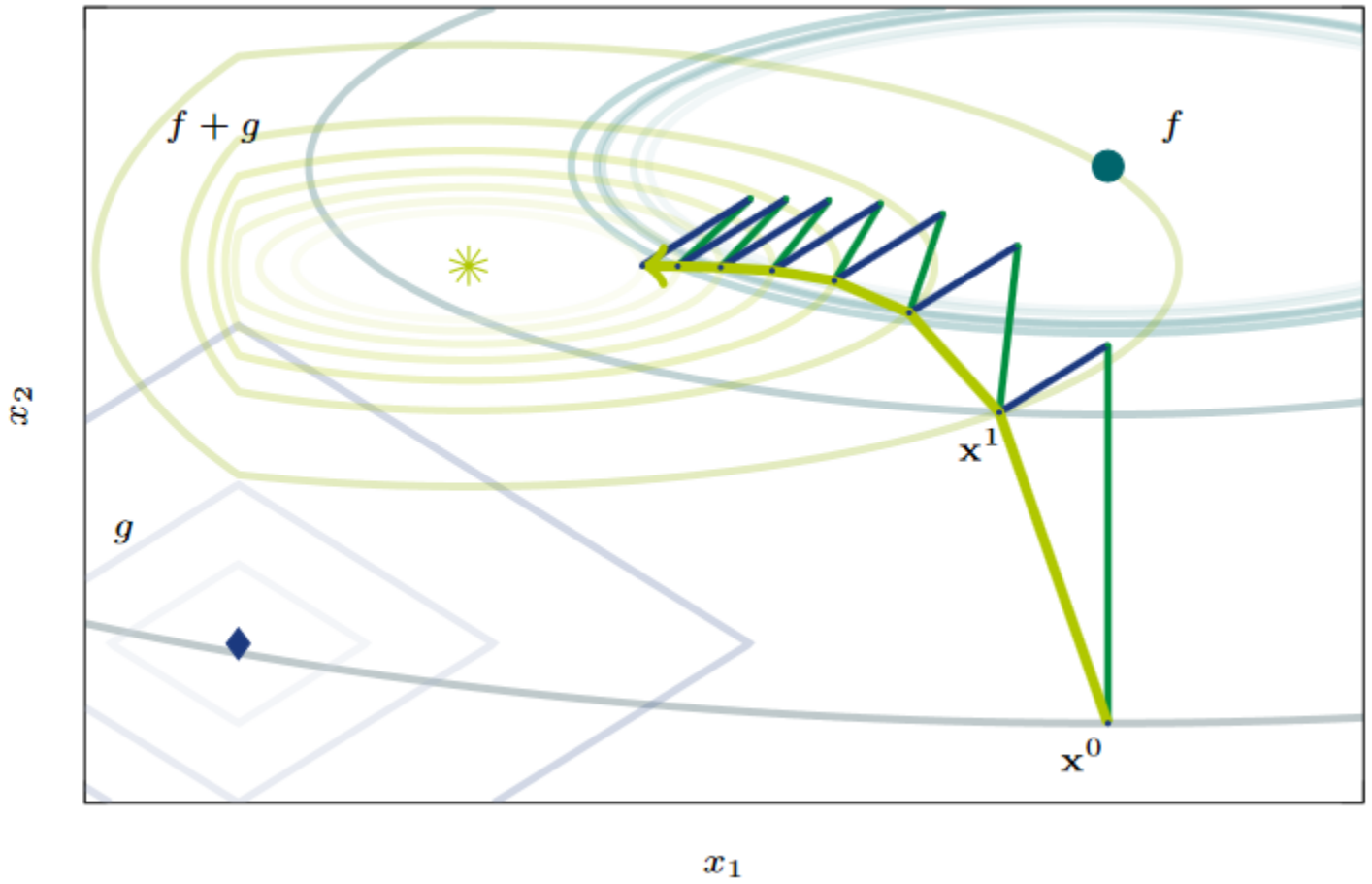
$\boldsymbol{x}^0 \in \mathcal{C}$
FOR $k = 0$ TO $+\infty$
$\quad \boldsymbol{x}^{k+1} = \Pi_{\mathcal{C}}\left(\boldsymbol{x}^k - \gamma \nabla f\left(\boldsymbol{x}^k\right)\right)$
END

- Έχουμε τον αλγόριθμο Proximal Gradient Descend

$\boldsymbol{x}^0 \in \mathbb{R}^d, \gamma \in (0, 2L_f^{-1}]$
FOR $k = 0$ TO $+\infty$
$\quad \boldsymbol{x}^{k+1} = prox_{\gamma g}\left(\boldsymbol{x}^k - \gamma \nabla f\left(\boldsymbol{x}^k\right)\right)$
END

Composite Convex Optimization

- ΠΑΡΑΔΕΙΓΜΑ



Composite Convex Optimization

Accelerated Proximal Gradient Αλγόριθμος

- Κατ' αναλογία με τον Accelerated Gradient Descend, όπου έχει εφαρμοστεί η μέθοδος επιτάχυνσης του Nesterov, μπορούμε να εφαρμόσουμε και εδώ μια παρόμοια τεχνική και να καταλήξουμε σε έναν αλγόριθμο ο οποίος είναι γνωστός με το όνομα Fast Proximal Gradient Algorithm

Algorithm 2 Fast proximal gradient algorithm (FPG) [12]

1: Set $\mathbf{v}^0 = \mathbf{x}^{-1} \in \mathbb{R}^n$, $\gamma \in (0, L_f^{-1}]$, and $\theta_0 = 1$
2: **for** $k = 0, 1, \ldots$ **do**
3: $\quad \mathbf{x}^k = \text{prox}_{\gamma g}(\mathbf{v}^k - \gamma \nabla f(\mathbf{v}^k))$
4: $\quad \theta_{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4\theta_k^2}\right)$
5: $\quad \mathbf{v}^{k+1} = \mathbf{x}^k + (\theta_k - 1)\theta_{k+1}^{-1}(\mathbf{x}^k - \mathbf{x}^{k-1})$

Συνήθεις Όροι Κανονικοποίησης

(και οι αντίστοιχοι τελεστές εγγύτητας)

Συνήθεις Όροι Κανονικοποίησης

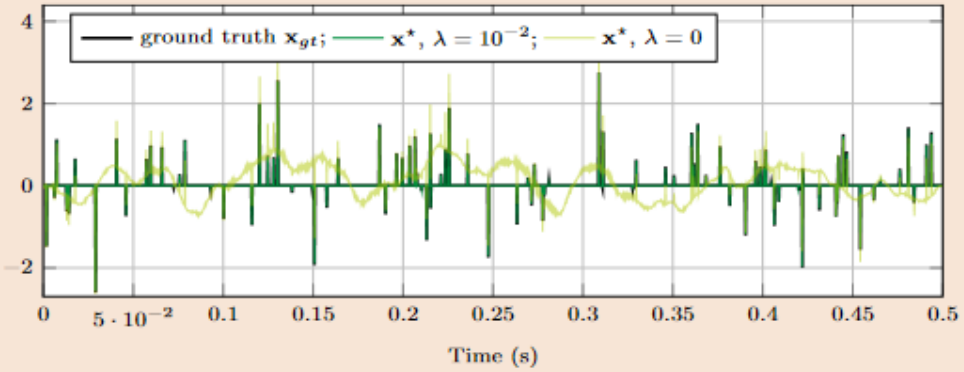
g	$\text{prox}_{\gamma g}$	**Properties**
$\|\mathbf{x}\|_0$	x_i if $\|x_i\| > \sqrt{2\gamma}$, 0 elsewhere	nonconvex, separable
$\|\mathbf{x}\|_1$	$\mathcal{P}_+(\mathbf{x}-\gamma) - \mathcal{P}_+(-\mathbf{x}-\gamma)$	convex, separable
$\|\mathbf{x}\|$	$\max\{0, 1-\gamma/\|\mathbf{x}\|\}\mathbf{x}$	convex
$\|\mathbf{X}\|_*$	$\mathbf{U}\,\text{diag}\,(\mathcal{P}_+(\boldsymbol{\sigma}-\gamma))\,\mathbf{V}^{\mathsf{H}}$	convex
$\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$	$(\mathbf{A}^{\mathsf{H}}\mathbf{A}+\gamma^{-1}\mathsf{Id})^{-1}(\mathbf{A}^{\mathsf{H}}\mathbf{b}+\gamma^{-1}\mathbf{x})$	convex
$\mathscr{S}$	$\Pi_{\mathscr{S}}$	
$\{\mathbf{x} \mid \|\mathbf{x}\|_0 \leq m\}$	$\mathcal{P}_m\mathbf{x}$	nonconvex
$\{\mathbf{x} \mid \|\mathbf{x}\| \leq r\}$	$r/\|\mathbf{x}\|\mathbf{x}$ if $\|\mathbf{x}\| > r$, $\mathbf{x}$ otherwise	convex
$\{\mathbf{x} \mid l \leq \mathbf{x} \leq u\}$	$\min\{u, \max\{l, x_i\}\}\ \forall\ i = 1, \ldots, n$	convex, separable
$\{\mathbf{X} \mid \text{rank}(\mathbf{X}) \leq m\}$	$\mathbf{U}\,\text{diag}(\mathcal{P}_m\boldsymbol{\sigma})\mathbf{V}^{\mathsf{H}}$	nonconvex
$\{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}\}$	$\mathbf{x} + \mathbf{A}^{\mathsf{H}}(\mathbf{A}\mathbf{A}^{\mathsf{H}})^{-1}(\mathbf{b}-\mathbf{A}\mathbf{x})$	convex

Table 1: Table showing the proximal operators of a selection of functions g and indicator function $\delta_{\mathscr{S}}$ with sets $\mathscr{S}$. Here given a n long vector $\mathbf{x}$, $\mathcal{P}_+\mathbf{x}$ returns $[\max\{0, x_1\}, \ldots, \max\{0, x_n\}]^{\mathsf{T}}$ while $\mathcal{P}_m\mathbf{x}$ returns a copy of $\mathbf{x}$ with all elements set to 0 except for the m largest in modulus. The matrices $\mathbf{U}$ and $\mathbf{V}$ are the result of a SVD: $\mathbf{X} = \mathbf{U}\,\text{diag}(\boldsymbol{\sigma})\mathbf{V}^{\mathsf{H}}$ where $\boldsymbol{\sigma}$ is the vector containing the singular values of $\mathbf{X}$. See [47, Sec. 6.9] for a more exhaustive list of proximal operators.

Εφαρμογές

Εφαρμογές: Αραιά Inverse Προβλήματα

Example 1 Sparse Deconvolution



Deconvolution seeks to recover the input signal $\mathbf{x}^\star$ from the available output signal $\mathbf{y}$ of a LTI system. A FIR $\mathbf{h}$ can be used to model the LTI system in terms of convolution. In a single-channel case with low SNR, deconvolution can easily become an ill-posed problem ($\lambda = 0$) and regularization must be added in order to achieve meaningful results. If $\mathbf{x}^\star$ is assumed to be sparse, a sparsity-inducing regularization function can be included in the optimization problem (LASSO):

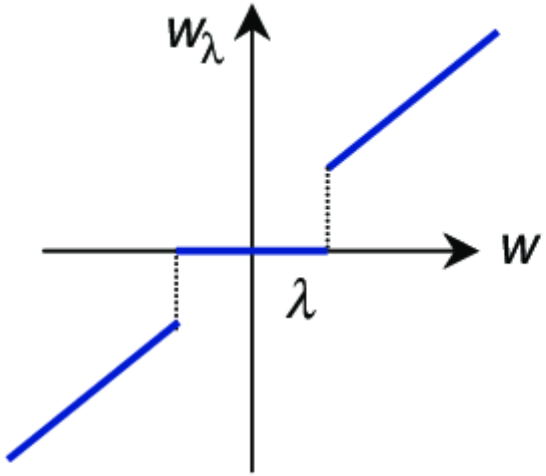
$$\mathbf{x}^\star = \underset{\mathbf{x}}{\operatorname{argmin}} \quad \underbrace{\tfrac{1}{2}\|\mathbf{h} * \mathbf{x} - \mathbf{y}\|^2}_{f(\mathbf{x})} + \underbrace{\lambda\|\mathbf{x}\|_1}_{g(\mathbf{x})}, \tag{1}$$

where $*$ indicates convolution and λ is a scalar that balances the weight between the regularization function and the data fidelity function.

StructuredOptimization code snippet:

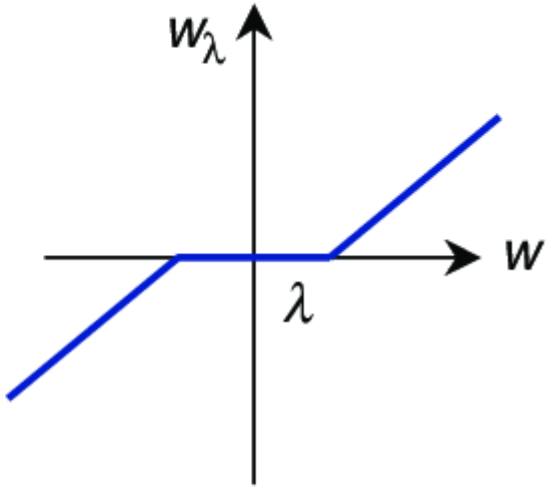
```
Fs = 4000 # sampling frequency
x = Variable(div(Fs,2)) # 'ls' short-hand
                        # for '0.5*norm(...)^2'
@minimize ls(conv(x,h)-y)+lambda*norm(x,1)
```

Αραιά Inverse Προβλήματα (ISTA/FISTA)



$$\delta_\lambda(w) = \begin{cases} w & |w| \geq \lambda \\ 0 & |w| < \lambda \end{cases}$$

(a) Hard thresholding

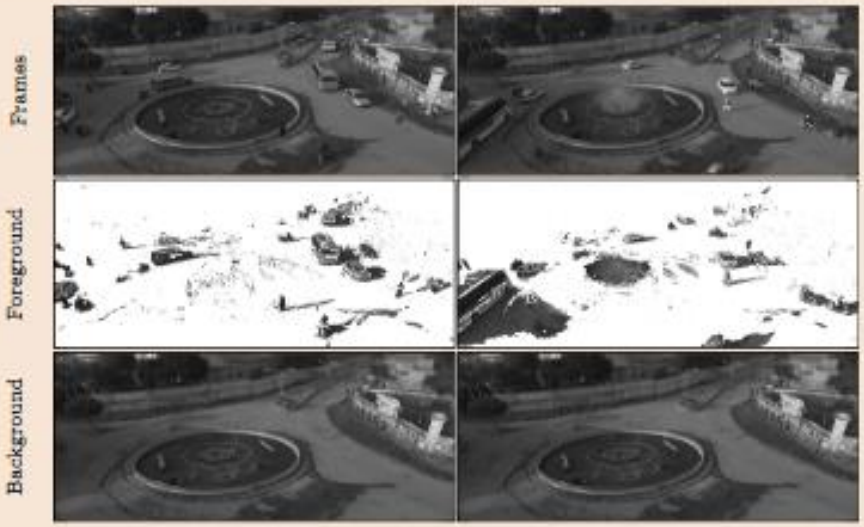


$$\delta_\lambda(w) = \begin{cases} \text{sgn}(w)(|w| - \lambda) & |w| \geq \lambda \\ 0 & |w| < \lambda \end{cases}$$

(b) Soft thresholding

Διαχωρισμός Υποβάθρου / Προσκηνίου

Example 4 Video background removal



The frames of a video can be viewed as a superposition of a moving foreground to a steady background. Splitting the background form the foreground can be a difficult task due to the continuous changes happening in different areas of the frames. The following optimization problem can be posed to deal with such task:

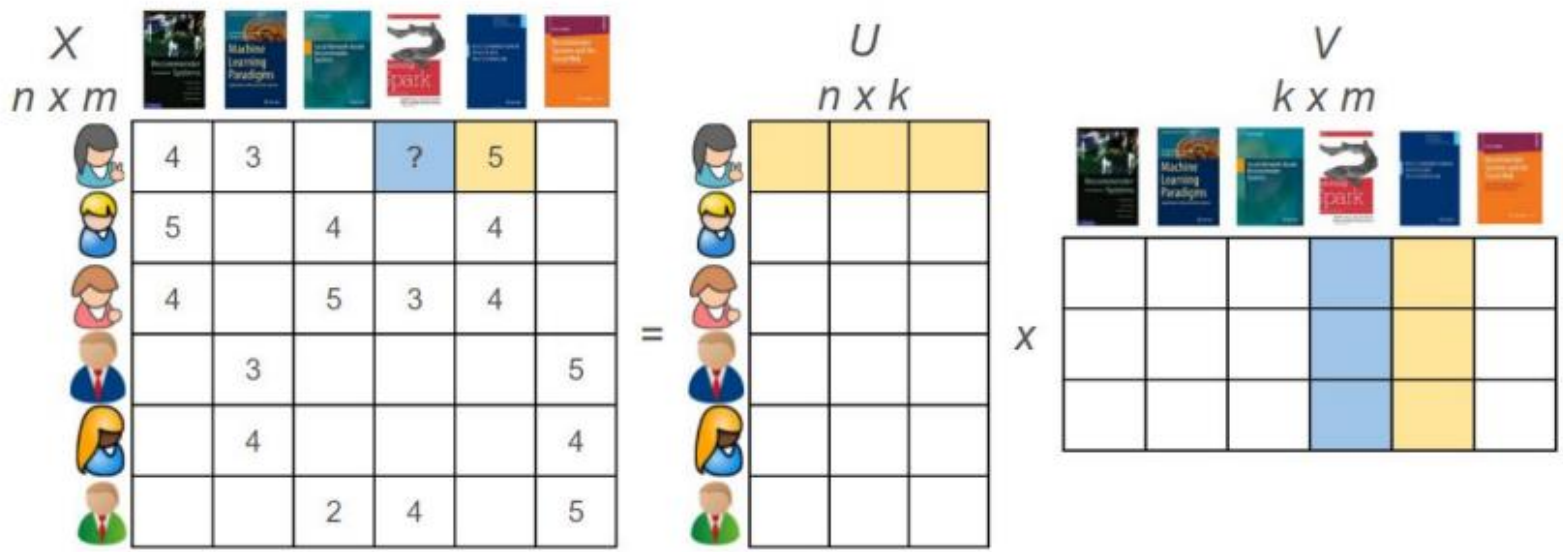
$$\underset{\mathbf{L},\mathbf{S}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{L}+\mathbf{S}-\mathbf{Y}\|^2+\lambda\|\text{vec}(\mathbf{S})\|_1 \quad \text{subject to} \quad \text{rank}(\mathbf{L}) \leq 1.$$

```
StructuredOptimization :
@minimize ls(L+S-Y)+
                lambda*norm(S,1)
st rank(L) <= 1
```

Here $\mathbf{Y} \in \mathbb{R}^{nm \times l}$ consists of a matrix in which the l-th column contains the pixel values of the vectorized l-th frame with dimensions $n \times m$. The optimization problem, also known as robust PCA, decomposes $\mathbf{Y}$ into a sparse matrix $\mathbf{S}$, representing the foreground changes and a rank-1 matrix $\mathbf{L}$ consisting of the constant background, whose columns are linearly dependent.

Εφαρμογές: Συμπλήρωση Πίνακα

- The Netflix problem



Εφαρμογές: Συμπλήρωση Πίνακα (SVT)

- Ορίζουμε

$$C = \{X \in \mathbb{R}^{m \times n} \text{ s.t. } X(i,j) = M(i,j) \quad \forall (i,j) \in \Omega\}$$

ή εναλλακτικά

$$C = \{X \in \mathbb{R}^{m \times n} \text{ s.t. } |X(i,j) - M(i,j)| \leq \varepsilon \quad \forall (i,j) \in \Omega\}$$

- Και λύνουμε το πρόβλημα

$$(P_3) \quad \min_{X \in C} \ \tau \|X\|_* + \frac{1}{2}\|X\|_F^2$$

g	$\text{prox}_{\gamma g}$
$\|\mathbf{x}\|_0$	x_i if $\|x_i\| > \sqrt{2\gamma}$, 0 elsewhere
$\|\mathbf{x}\|_1$	$\mathcal{P}_+(\mathbf{x} - \gamma) - \mathcal{P}_+(-\mathbf{x} - \gamma)$
$\|\mathbf{x}\|$	$\max\{0, 1 - \gamma/\|\mathbf{x}\|\}\mathbf{x}$
$\|\mathbf{X}\|_*$	$\mathbf{U}\,\text{diag}\,(\mathcal{P}_+(\boldsymbol{\sigma} - \gamma))\,\mathbf{V}^{\mathsf{H}}$
$\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$	$(\mathbf{A}^{\mathsf{H}}\mathbf{A} + \gamma^{-1}\text{Id})^{-1}(\mathbf{A}^{\mathsf{H}}\mathbf{b} + \gamma^{-1}\mathbf{x})$

Μείωση θορύβου με διατήρηση ακμών

Example 5 Total variation de-noising

ground truth $\mathbf{X}_{gt}$ noisy $\mathbf{Y}$ denoised $\mathbf{X}^\star$

```
StructuredOptimization

V = Variation(size(Y)); U = Variable(size(V,1)...)
@minimize ls(-V'*U+Y) + conj(lambda*norm(U,2,1,2))
X = Y-V'*(~U)
```

Total variation de-noising seeks to remove noise from a noisy image whose pixels are stored in the matrix $\mathbf{Y} \in \mathbb{R}^{n\times m}$. This technique relies on the assumption that neighbor pixels of the sought uncorrupted image $\mathbf{X}^\star$ should be similar, namely that $\sqrt{|x^\star_{i+1,j} - x^\star_{i,j}|^2 + |x^\star_{i,j+1} - x^\star_{i,j}|^2}$ should be small, where $x^\star_{i,j}$ is the (i,j)-th component of $\mathbf{X}^\star$. This enforces the image to have sharp edges, namely a sparse gradient. The following optimization problem can be formulated:

$$\text{(a)} \ \underset{\mathbf{X}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{X}-\mathbf{Y}\|^2 + \lambda\|\mathsf{V}\mathbf{X}\|_{2,1} \quad \text{(b)} \ \underset{\mathbf{U}}{\text{minimize}} \ \frac{1}{2}\|-\mathsf{V}^*\mathbf{U}+\mathbf{Y}\|^2 + g^*(\mathbf{U})$$

Here the operator $\mathsf{V} \in \mathscr{L}(\mathbb{R}^{n\times m}, \mathbb{R}^{nm\times 2})$ maps $\mathbf{X}$ into a matrix having in its j-th column the vectorized forward finite difference gradient over the j-th direction. The operator V appears in the nonsmooth part of the cost function $g(\cdot) = \lambda\|\cdot\|_{2,1}$ and leads to a non-trivial proximal operator. Here the mixed norm $\|\cdot\|_{2,1}$ consists of the sum of the l_2-norm of the rows of $\mathsf{V}\mathbf{X}$. Using Fenchel's duality theorem it is possible to convert the problem into (b) which can instead be solved efficiency using proximal gradient algorithms.