

Μέθοδοι βελτιστοποίησης

Μη λεια βελτιστοποίηση- *Sub – Gradient Schemes*

A. N. Γιαννακόπουλος

Ο.Π.Α

Εαρινό Εξάμηνο 2025-2026

1 Βασικές έννοιες

Το υποδιαφορικό (sub - gradient) και οι ιδιότητες του

Έστω $f : D \subset \mathbb{R}^d \rightarrow \mathbb{R}$ μια συνάρτηση d -μεταβλητών.

Ορισμός

Το διάνυσμα v λέμε ότι ανήκει στο υποδιαφορικό της f στο σημείο $x_0 \in D$ αν

$$f(x) \geq f(x_0) + \langle v, x - x_0 \rangle, \quad \forall x \in D.$$

Συμβολίζουμε $v \in \partial f(x_0)$.

Το υποδιαφορικό αποτελείται από τις κλίσεις όλων των υπερεπιπέδων που υποστηρίζουν την f στο σημείο x_0 .

Το υποδιαφορικό $\partial f(x_0) \subset \mathbb{R}^d$ είναι ένα σύνολο το οποίο μπορεί να περιέχει περισσότερα του ενός στοιχεία!

Είναι μονοσύνολο αν και μόνο αν η συναρτησης f είναι διαφορίσιμη στο x_0 και τότε $\partial f(x_0) = \{Df(x_0)\}$.

Το υποδιαφορικό υπάρχει πάντοτε για κυρτές συναρτήσεις (και όχι μόνο!) και είναι ένα κυρτό σύνολο με την σειρά του.

Παράδειγμα

Δίνεται η συνάρτηση $f(x) = |x|$ η οποία δεν είναι διαφορίσιμη στο $x_0 = 0$.

Οποιοδήποτε $v \in [-1, 1]$ ανήκει στο υποδιαφορικό της f ,
δηλ. $\partial f(0) = [-1, 1]$.

Παράδειγμα

Δίνεται η συναρτηση f , που ορίζεται $x = (x_1, \dots, x_d) \mapsto \sum_{i=1}^d |x_i|$ που δίνει την ℓ_1 νόρμα ενός διανύσματος του $\mathbb{R}^d$.

Η συνάρτηση αυτή είναι μη διαφορίσιμη στο $x = (0, \dots, 0)$ και μπορεί κανείς να δει απο τον ορισμό ότι

$$\partial f(x) = \{v \in \mathbb{R}^d : v_i = \operatorname{sgn}(x_i) \text{ αν } x_i \neq 0 \text{ και } v_i \in [-1, 1] \text{ αν } x_i = 0\}$$

Και γιατί να με ενδιαφέρει;

Πολλά απο τα κλασσικά υποδείγματα της μηχανικής μαθησης χρησιμοποιούν μη λείες (μη διαφορισιμες) συναρτησεις κανονικοποίησης.

Το πιο χαρακτηριστικό παράδειγμα είναι το μοντέλο Lasso για την γραμμική παλινδρομηση

$$\min_w J(w) = \min_w \|Xw - y\|^2 + \lambda \|w\|_{\ell_1}$$

το οποίο προσπαθεί να προσαρμόσει στα δεδομένα ένα γραμμικό μοντέλο $y = \langle w, x \rangle$, με την προϋπόθεση να χρησιμοποιηθούν όσο το δυνατόν λιγότερες απο τις μεταβλητές $x = (x_1, \dots, x_d)$ στο υπόδειγμα (δηλ. να βρεθεί το πιο οικονομικό μοντέλο απο πλευράς διαστάσεων).

Η συνάρτηση σφάλματος J που σχετίζεται με το μοντέλο Lasso είναι μη παραγωγισιμη σε οποιοδήποτε w το οποίο έχει κάποια συνιστώσα $w_i = 0$.

Κατά συνέπεια, ένα σχήμα βαθμίδας για την επίλυση του προβλήματος βελτιστοποίησης (και ισοδύναμα την εύρεση του κατάλληλου μοντέλου) δεν μπορεί να χρησιμοποιηθεί.

Τι να κανουμε;

Το υποδιαφορικό μπορεί να υποκαταστήσει την βαθμίδα για τα προβλήματα που δεν είναι λεία.

Θεώρημα

Έστω $f : U \subset \mathbb{R}^d \rightarrow \mathbb{R}$ κυρτή συναρτηση.

Το σημείο $x_0 \in U$ είναι τοπικό ελάχιστο της f αν και μόνο αν

$$0 \in \partial f(x_0)$$

Παρατηρείστε την ομοιότητα με την περίπτωση όπου η f είναι διαφορίσιμη στην οποία ισχυει ότι

$$f(x_0) = 0$$

Με βάση το παραπάνω μπορούμε να τροποποιήσουμε τον αλγόριθμο στον subgradient algorithm

- Καθορίζουμε μια αρχική θέση x_0 και μια ακρίβεια $\epsilon > 0$.
- Επιλέγουμε κάθε φορά ένα $v_k \in \partial f(x_k)$.
- Ανανεώνουμε

$$x_{k+1} = x_k - \alpha_k v_k$$

- Όταν $\|x_{k+1} - x_k\| < \epsilon$ σταματάμε και επιστρέφουμε $x = x_{k+1}$.

Το x το οποίο επιστρέφουμε είναι μια προσέγγιση του ελαχιστου δεδομένου ότι για το x αυτό ισχύει

$$v \simeq 0, \text{ και } v \in \partial f(x)$$

οπότε $x \in \arg \min f(x)$.

Αυτό είναι όλο;

Ο παραπάνω αλγόριθμος είναι ενδιαφέρων όμως παρότι μοιάζει οπτικά με το gradient descent δεν ισχύει πάντοτε ότι ένα αρνητικό υποδιαφορικό είναι απαραίτητα μια κατεύθυνση μείωσης της τιμής της f .

Παράδειγμα

Έστω $f(x) = |x_1| + |x_2|$ και $x_0 = (1, 0)$.

$v = (1, w) \in \partial f(x_0)$ για οποιοδήποτε $w \in [-1, 1]$ όμως η $-v$ για καποιες επιλογές του w μπορεί να μην είναι κατευθυνση μειωσης της τιμής της f .

Κατά συνέπεια η ακολουθία $\{f(x_k)\}$ δεν είναι απαραίτητα μονότονη (φθίνουσα).

Πιο γενικά:

Απο τον ορισμό του $v_k \in \partial f(x_k)$ έχουμε **μόνο** ότι

$$\begin{aligned} f(x_{k+1}) &= f(x_k - \alpha_k v_k) \geq f(x_k) + \langle v_k, x_k - \alpha_k v_k - x_k \rangle \\ &= f(x_k) - \alpha_k \langle v_k, v_k \rangle = f(x_k) - \alpha_k \|v_k\|^2 \end{aligned}$$

Η σχέση αυτή δεν μας εξασφαλίζει ότι $f(x_{k+1}) \leq f(x_k)$, για την ακρίβεια θα μπορούσε $f(x_{k+1}) = f(x_k) + c$ για κάποιο $c > 0$.

Για να ξεφύγουμε απο το παραπάνω:

Σε κάθε βήμα της μεθόδου υπολογίζουμε το

$$f_k^* = \min\{f(x_0), f(x_1), \dots, f(x_k)\},$$

και παίρνουμε την ακολουθία $\{f^*(x_k)\}$.

Παρατηρείστε ότι η ακολουθία $\{f^*(x_k)\}$ μπορεί να παραχθεί και απο τον επαναληπτικό τύπο

$$f_k^* = \min\{f_{k-1}^*, f(x_k)\},$$

και είναι μια φθίνουσα ακολουθία.

Σε κάθε βήμα k λοιπόν κρατάμε την πιο μικρή τιμή απο αυτές που έχει πετυχει η συνάρτηση μέχρι τώρα και σαν x_k την θέση στην οποία την πέτυχε,

$$f(x_k^*) = \min\{f(x_0), \dots, f(x_k)\}$$

και χρησιμοποιούμε τις ακολουθίες $\{x_k^*, f_k^*\}$.

Τα βήματα a_k επιλέγονται συνήθως έτσι ώστε να ικανοποιούν κάποια απο τις παρακατω ιδιότητες:

- $a_k = h$ σταθερό βήμα.
- $a_k = h/\|v_k\|_2$
- $\sum_k a_k^2 < \infty$ αλλά $\sum_k a_k = \infty$,
- $\lim_k a_k = 0$ αλλά $\sum_k a_k = \infty$

Σημειώνεται οτι στην μέθοδο subgradient τα μεγέθη των βημάτων καθορίζονται απο την αρχή και δεν ανανεώνονται σε κάθε βημα ανάλογα με την θέση x_k που βρεθήκαμε.

Στις δύο τελευταίες περιπτώσεις μπορεί να αποδειχθεί ότι η ακολουθία $f^*(x_k) \rightarrow f^*$ όπου f^* είναι η ελάχιστη τιμή της f .

Subgradient algorithm

- ❶ Καθορίζουμε την ακρίβεια ϵ , μια ακολουθία βημάτων $\{\alpha_k\}$ όπως παραπάνω.
- ❷ Επιλέγουμε ένα αρχικό σημείο x_0 , θέτουμε $k = 0$, $x_* = x_0$, επιλέγουμε $v_0 \in \partial f(x_0)$ και υπολογίζουμε $x_1 = x_0 - \alpha_0 v_0$.
- ❸ Στο k βήμα:
 - ❶ Αν $\|v_k\| < \epsilon$ ή $k + 1 > k_{max}$ σταματάμε και επιστρέφουμε το x_* .
 - ❷ Θέτουμε $x_{k+1} = x_k - \alpha_k v_k$ και επιλέγουμε ένα $v_{k+1} \in \partial f(x_{k+1})$.
 - ❸ Θέτουμε $k := k + 1$.
 - ❹ Αν $f(x_k) < f(x_*)$ θέτουμε $x_* = x_k$
 - ❺ Επιστρέφουμε στο βήμα 3.1.
- ❹ Ο αλγόριθμος επιστρέφει το x_* .

Γιατί δουλεύει;

Εφόσον τα $-v_k$, $v_k \in \partial f(x_k)$ δεν αποτελούν απαραίτητα διευθύνσεις μείωσης της τιμής της συνάρτησης δεν είναι απαραίτητο ότι η x_{k+1} είναι καλύτερη από την x_k όσον αφορά την τιμή της συνάρτησης (όπως συμβαίνει στους αλγορίθμους βαθμίδας).

Όμως, η ακολουθία που παράγει η subgradient μέθοδος έχει την ακόλουθη σημαντική ιδιότητα.

Λήμμα

Αν x^* ελάχιστο της κυρτής συναρτησης f , τότε η ακολουθία $\{x_k\}$, με $x_{k+1} = x_k - \alpha_k v_k$, $v_k \in \partial f(x_k)$ έχει την ιδιότητα

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\|,$$

αρκεί

$$0 < \alpha_k < \frac{2}{\|v_k\|^2} (f(x_k) - f(x^*)).$$

Απόδειξη

Παρατηρούμε ότι

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - \alpha_k v_k \implies \\ \|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 + \alpha_k^2 \|v_k\|^2 + 2\alpha_k \langle v_k, x^* - x_k \rangle \end{aligned}$$

Απο τον ορισμό του υποδιαφορικού $v_k \in \partial f(x_k)$, εφαρμοσμένο στο $x = x^*$ έχουμε

$$\begin{aligned} f(x^*) &\geq f(x_k) + \langle v_k, x^* - x_k \rangle \implies \\ \langle v_k, x^* - x_k \rangle &\leq f(x^*) - f(x_k) \leq 0 \end{aligned}$$

Συνεπώς, η πρώτη σχέση γίνεται

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \alpha_k^2 \|v_k\|^2 + 2\alpha_k (f(x^*) - f(x_k)),$$

απο όπου για α_k επαρκώς μικρό προκύπτει το αποτέλεσμα.

Ακολουθίες με την ιδιότητα $\|x_{k+1} - x^*\| \leq \|x_k - x^*\|$ ονομάζονται ακολουθίες Fejer .

Μία ακολουθία Fejer συγκλίνει αλλά δεν είναι σίγουρο οτι συγκλίνει στο x^* !

Αν $\sum_k \alpha_k^2 < \infty$ και $\sum_k \alpha_k = \infty$ τότε μπορεί να δείξει κανείς οτι η ακολουθία $\{x_k\}$ συγκλίνει στο x^* με υπογραμμικό ρυθμό.

Άλλες επιλογές των α_k μπορεί να οδηγήσουν σε καλύτερες ιδιότητες συγκλισης.

Οι μέθοδοι subgradient παρουσιάζουν προβλήματα όπως π.χ. η αδυναμία να καθορίσουμε ένα κριτήριο τερματισμού, η αργή σύγκλιση, η περίπτωση τα sugradient να μην είναι διευθύνσεις κατάβασης κλπ.

Παρόλα αυτά είναι πολυ δημοφιλείς στην πράξη λόγω της απλότητας τους και χρησιμοποιούνται αρκετά.

Εφαρμογή lasso

$$\min_w \frac{1}{2} \|\mathbf{D}w - y\|_2^2 + \lambda \|w\|_1$$

$$w_{k+1} = w_k - \alpha_k \mathcal{E}_k - \alpha v_k,$$

$$v_{k,i} = \begin{cases} 1 & w_{k,i} > 0 \\ -1 & w_{k,i} < 0 \\ u \in [-1, 1] & w_{k,i} = 0 \end{cases}$$

Μπορούμε να επιλέξουμε π.χ. $u = 0$, αλλά σε κάθε περίπτωση πρέπει να ελεγχουμε το αν η συναρτηση παιρνει μικροτερες τιμες στην νεα διευθυνση

Επεκτάσεις

Πολλες φορές δεν είναι τόσο εύκολο να υπολογίσουμε ένα στοιχείο στο υποδιαφορικό $\partial f(x_k)$.

Για τον λόγο αυτό πολλές φορές καταφευγουμε σε προσεγγίσεις, ή σε ομαλοποιήσεις του υποδιαφορικού.

Μία τέτοια παραλλαγή είναι το proximal point algorithm

Proximal point algorithm

Ορισμός

Έστω $f : \mathbb{R}^d \rightarrow \mathbb{R}$, κυρτή και κάτω ημισυνεχής συνάρτηση.

Ο τελεστής $\text{prox}_f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ που ορίζεται από το

$$x \mapsto x^* = \arg \min_{z \in \mathbb{R}^d} \left(\frac{1}{2} \|x - z\|_2^2 + f(z) \right) =: \text{prox}_f(x)$$

ονομάζεται τελεστής συνάφειας (*proximity operator*)

- Είναι μονοσήμαντα ορισμένος και αποτελεί γενίκευση του τελεστή προβολής.
- Είναι ψευδοσυστολή, δηλαδή

$$\|\text{prox}_f(x_1) - \text{prox}_f(x_2)\| \leq \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^d.$$

- Συνδέεται με το υποδιαφορικό της f ,

$$x^* = \text{prox}_f(x) \iff x - x^* \in \partial f(x^*)$$

και μπορούμε να γράψουμε $x^* = (I + \partial f)^{-1}(x)$.

Παράδειγμα

Αν $f : \mathbb{R} \rightarrow \mathbb{R}$ είναι η $f(x) = |x|$ τότε

$$\text{prox}_{\lambda f}(x) = \begin{cases} x - \lambda & \text{αν } x > \lambda, \\ 0 & \text{αν } -\lambda \leq x \leq \lambda, \\ x + \lambda & \text{αν } x < -\lambda. \end{cases}$$

Παράδειγμα (Soft threshold or shrinkage operator)

Αν $f : \mathbb{R}^d \rightarrow \mathbb{R}$ είναι η $f(x) = \|x\|_1$ τότε

$$\text{prox}_{\lambda f}(x) = ([\text{prox}_{\lambda f}(x)]_1, \dots, [\text{prox}_{\lambda f}(x)]_d),$$

με

$$[\text{prox}_{\lambda f}(x)]_i = \begin{cases} x_i - \lambda & \text{αν } x_i > \lambda, \\ 0 & \text{αν } -\lambda \leq x_i \leq \lambda, \\ x_i + \lambda & \text{αν } x_i < -\lambda. \end{cases}$$

Προσέγγιση Moreau - Yosida

Η προσέγγιση Moreau-Yosida αποτελεί μια διαδικασία λείας προσέγγισης για μια μη λεία συνάρτηση f .

Ορισμός

Η οικογένεια συναρτησεων $\{f_\lambda : \lambda > 0\}$, $f_\lambda : \mathbb{R}^d \rightarrow \mathbb{R}$, που ορίζεται ως

$$f_\lambda(x) := \min_{z \in \mathbb{R}^d} \left(\frac{1}{2\lambda} \|x - z\|_2^2 + f(z) \right)$$

ονομάζεται προσέγγιση Moreau-Yosida της f .

Η προσέγγιση M-Y μπορεί να εκφραστεί και συναρτήσει του proximity operator ως

$$f_\lambda(x) = \frac{1}{2\lambda} \|x - \text{prox}_{\lambda f}(x)\|_2^2 + f(\text{prox}_{\lambda f}(x))$$

Ιδιότητες της προσέγγισης Moreau-Yosida

- Η f_λ είναι διαφορίσιμη με συνεχή κατά Lipschitz παράγωγο Df_λ , και

$$Df_\lambda(x) = \frac{1}{\lambda}(x - \text{prox}_{\lambda f}(x)).$$

- Η f_λ έχει τους ίδιους ελαχιστοποιητές με την f ,

$$x^* \in \arg \min_{z \in \mathbb{R}^d} f(z) \iff x^* \in \arg \min_{z \in \mathbb{R}^d} f_\lambda(z)$$

- $\lim_{\lambda \rightarrow 0^+} f_\lambda(x) = f(x)$.

Παράδειγμα

Αν $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = |x|$, τότε

$$f_{\lambda}(x) = \begin{cases} x - \frac{\lambda}{2} & \text{αν } x > \lambda, \\ \frac{x^2}{2\lambda} & \text{αν } -\lambda \leq x \leq \lambda, \\ -x - \frac{\lambda}{2} & \text{αν } x < -\lambda. \end{cases}$$

Proximal point algorithm

Ο αλγόριθμος proximal point βασίζεται στο επαναληπτικό σχήμα

$$x_{n+1} = \text{prox}_{\lambda f}(x_n)$$

Η ακολουθία $\{x_n\}$ ονομάζεται proximal ακολουθία.

Για ένα σταθερό σημείο του παραπάνω σχήματος ισχύει ότι

$$x = \text{prox}_{\lambda f}(x) \iff x \in \arg \min_{z \in \mathbb{R}^d} f(z)$$

Ποιά θα είναι η επιλογή του λ ;

- Το λ δεν μπορεί να είναι πολύ μικρό γιατί $\lim_{\lambda \rightarrow 0^+} \text{prox}_{\lambda f}(x) = x$ (εξαιρετικά αργή συγκλιση).
- Το λ δεν μπορεί να είναι πολύ μεγάλο, γιατί διαταράσσεται ο τοπικός χαρακτήρας της συνθήκης πρώτης τάξης.

Σχόλιο

Εφόσον $\text{prox}_{\lambda f}(x) = (I + \lambda \partial f)^{-1}(x)$ βλέπουμε πως η proximal sequence ισοδύναμα μπορεί να οριστεί ως

$$x_{n+1} = (I + \lambda \partial f)^{-1}(x_n) \iff x_{n+1} + \lambda \partial f(x_{n+1}) = x_n.$$

Στην περίπτωση που η f είναι λεία, η μέθοδος αυτή μας δίνει

$$x_{n+1} = x_n - \lambda Df(x_{n+1})$$

που μπορεί να την θεωρήσει κανείς σαν μια πεπλεγμένη μέθοδο βαθμίδας (implicit gradient method).

Αλγόριθμος (Proximal point)

- ① Επιλέγουμε μια ακολουθία $\{\lambda_n\} \subset \mathbb{R}_+$ τέτοια ώστε $\sum_n \lambda_n = \infty$ και κάποια ακρίβεια $\epsilon > 0$.
- ② Για ένα κατάλληλο x_1 ακολουθούμε την επαναληπτική διαδικασία

$$x_{k+1} = \text{prox}_{\lambda_n f}(x_n) := \arg \min_{z \in \mathbb{R}^d} \left(f(z) + \frac{1}{2\lambda_n} \|x - z\|_2^2 \right)$$

μέχρις ότου το κριτήριο σύγκλισης ικανοποιηθεί.

Η ακολουθία $\{x_n\}$ είναι μια ελαχιστοποιητική ακολουθία που έχει την ιδιότητα Fejer και συγκλίνει σε ένα σημείο που ανήκει στο σύνολο των ελαχιστοποιητών της f .

Αλγόριθμος (Inertial proximal point (relaxed version))

- ① Επιλέγουμε μια ακολουθία $\{\lambda_n\} \subset \mathbb{R}_+$ τέτοια ώστε $\sum_n \lambda_n = \infty$, κάποια ακρίβεια $\epsilon > 0$ και $\theta \in (-1, 1)$.
- ② Για ένα κατάλληλο x_1 ακολουθούμε την επαναληπτική διαδικασία

$$x_{k+1} = (1 + \theta) \text{prox}_{\lambda_n f}(x_n) - \theta x_n$$

μέχρις ότου το κριτήριο σύγκλισης ικανοποιηθεί.

Forward-backward αλγόριθμος

Ο αλγόριθμος αυτός χρησιμεύει για την εύρεση ελαχίστου σε προβλήματα

$$\min_{x \in \mathbb{R}^d} f_1(x) + f_2(x)$$

όπου

- f_1, f_2 κυρτές και κάτω ημισυνεχείς συναρτήσεις
- f_1 διαφορίσιμη με Lipschitz συνεχή παράγωγο
- f_2 μη λεία.

Ο αλγόριθμος forward-backward βασίζεται στην παρατήρηση ότι αν $x^* \in \arg \min_{z \in \mathbb{R}^d} (f_1(z) + f_2(z))$, οπότε η συνθήκη πρώτης τάξης δίνει

$$0 \in \partial(f_1 + f_2)(x^*) = Df_1(x^*) + \partial f_2(x^*)$$

Την συνθήκη αυτή την γράφουμε ως

$$\begin{aligned} (x^* - z) + (z - x^*) &\in Df_1(x^*) + \partial f_2(x^*) \implies \\ x^* - z &= \alpha Df_1(x^*), \text{ και } z - x^* \in \alpha \partial f_2(x^*) \implies \\ z &= x^* - \alpha Df_1(x^*), \text{ και } z = x^* + \alpha \partial f_2(x^*) \implies \\ z &= x^* - \alpha Df_1(x^*), \text{ και } x^* = (I + \alpha \partial)^{-1} f_2(z) = \text{prox}_{\alpha f}(z) \end{aligned}$$

Χρησιμοποιώντας λοιπόν την βοηθητική μεταβλητή z σπάμε το πρόβλημα σε 2:

$$z = x^* - \alpha Df_1(x^*), \text{ smooth: steepest descent}$$

$$x^* = \text{prox}_{\alpha f}(z), \text{ nonsmooth: proximal point}$$

Το παραπάνω το γράφουμε σαν ένα (επαναληπτικό) σχήμα σταθερού σημείου,

$$z_n = x_n - \alpha Df_1(x_n), \text{ smooth: steepest descent}$$

$$x_{n+1} = \text{prox}_{\alpha f}(z_n), \text{ nonsmooth: proximal point}$$

Για την επίτευξη καλύτερων ιδιοτήτων σύγκλισης μπορούμε να πάρουμε και το σχήμα χαλάρωσης (relaxed scheme)

$$\begin{aligned} z_n &= x_n - \alpha Df_1(x_n), \\ x_{n+1} &= x_n + \lambda_n (\text{prox}_{\alpha f}(z_n) - x_n) \\ &= (1 - \lambda_n)x_n + \lambda_n \text{prox}_{\alpha f}(z_n), \end{aligned}$$

Αλγόριθμος (Forward-backward)

- ❶ Επιλέγουμε $\alpha > 0$ επαρκώς μικρό, λ_n τέτοια ώστε $\sum_n \lambda_n(\beta - \lambda_n) = \infty$ για κάποιο $\beta > 1/2$ και κάποια ακρίβεια ϵ .
- ❷ Απο ένα κατάλληλο αρχικό σημείο (z_1, x_1) ακολουθούμε το επαναληπτικό σχήμα

$$\begin{aligned} z_n &= x_n - \alpha Df_1(x_n), \\ x_{n+1} &= (1 - \lambda_n)x_n + \lambda_n \text{prox}_{\alpha f}(z_n), \end{aligned}$$

εως ότου ικανοποιηθεί κάποιο κριτήριο σύγκλισης.

Για $\alpha < \frac{2}{L}$ όπου L η σταθερά Lipschitz της f_1 και $\{\lambda_n\}$ όπως παραπάνω η μέθοδος forward-backward συγκλίνει σε κάποιο ελαχιστοποιητή της $f_1 + f_2$.

Douglas-Rachford

Ο αλγόριθμος αυτός χρησιμεύει για την εύρεση ελαχίστου σε προβλήματα

$$\min_{x \in \mathbb{R}^d} f_1(x) + f_2(x)$$

όπου

- f_1, f_2 κυρτές και κάτω ημισυνεχείς συναρτήσεις
- f_1 μη λεία
- f_2 μη λεία.

Ο αλγόριθμος forward-backward βασίζεται στην παρατήρηση ότι αν $x^* \in \arg \min_{z \in \mathbb{R}^d} (f_1(z) + f_2(z))$, οπότε η συνθήκη πρώτης τάξης δίνει

$$0 \in \partial(f_1 + f_2)(x^*) = Df_1(x^*) + \partial f_2(x^*)$$

Την συνθήκη αυτή την γράφουμε ως

$$\begin{aligned} (2x^* - z - x^*) + (z - x^*) &\in \partial f_1(x^*) + \partial f_2(x^*) \implies \\ 2x^* - z &= x^* + \alpha \partial f_1(x^*), \text{ και } z \in x^* + \alpha \partial f_2(x^*) \implies \\ x^* &= (I + \alpha \partial f_1)^{-1}(2x^* - z), \text{ και } x^* = (I + \alpha \partial f_2)(z) \implies \\ x^* &= \text{prox}_{\alpha f_1}(2x^* - z), \text{ και } x^* = \text{prox}_{\alpha f_2}(z) \end{aligned}$$

Χρησιμοποιώντας λοιπόν την βοηθητική μεταβλητή z σπάμε το πρόβλημα σε 2 συνθήκες συμβατότητας

$$\begin{aligned} x^* &= \text{prox}_{\alpha f_2}(z), \\ x^* &= \text{prox}_{\alpha f_1}(2x^* - z) \end{aligned}$$

Αν απαλείψουμε το x^* οι παραπάνω σχέσεις δίνουν

$$\begin{aligned} 0 &= \text{prox}_{\alpha f_1}(2\text{prox}_{\alpha f_2}(z) - z) - \text{prox}_{\alpha f_2}(z) \implies \\ z &= z + \text{prox}_{\alpha f_1}(2\text{prox}_{\alpha f_2}(z) - z) - \text{prox}_{\alpha f_2}(z) =: G(z) \end{aligned}$$

την οποία προσεγγίζουμε με το σχήμα σταθερού σημείου

$$z_{n+1} = G(z_n)$$

ή το σχήμα χαλάρωσης (relaxed scheme)

$$z_{n+1} = z_n + \lambda_n(G(z_n) - z_n)$$

Επαναφέροντας την μεταβλητή $x_n = \text{prox}_{\alpha f_2}(z_n)$, παίρνουμε την ισοδύναμη μορφή

$$\begin{aligned} x_n &= \text{prox}_{\alpha f_2}(z_n) \\ z_{n+1} &= z_n \lambda_n(\text{prox}_{\alpha f_1}(2x_n - z_n) - x_n) \end{aligned}$$

Αλγόριθμος (Douglas-Ratchford)

- 1 Επιλέγουμε $\alpha > 0$, $\{\lambda_n\} \subset (0, 2]$, με $\sum_n \lambda_n(2 - \lambda_n) = \infty$ και μια ακριβεία $\epsilon > 0$.
- 2 Για κατάλληλη αρχική συνθήκη (x_1, z_1) ακολουθούμε το επαναληπτικό σχήμα

$$\begin{aligned}x_n &= \text{prox}_{\alpha f_2}(z_n) \\z_{n+1} &= z_n \lambda_n (\text{prox}_{\alpha f_1}(2x_n - z_n) - x_n)\end{aligned}$$

Το σχήμα Douglas-Ratchford συγκλίνει σε ένα ελαχιστοποιητή της $f_1 + f_2$.

Μέθοδοι δέσμης (bundle methods)

Στις μεθόδους δέσμης (bundle methods) η βασική ιδέα είναι να προσεγγίσουμε ολόκληρο το υποδιαφορικό (χρησιμοποιώντας στοιχεία του - subgradients - απο προηγούμενα βήματα της μεθόδου) και όχι να χρησιμοποιούμε σε κάθε επανάληψη της μεθόδου ένα αυθαίρετο στοιχείο του υποδιαφορικού.

Σε κάθε βήμα της μεθόδου k , έχουμε

- Το τρέχων σημείο των επαναλήψεων x_k .
- Τα δοκιμαστικά σημεία $z_1, z_2, \dots, z_{k-1}$ που έχουμε συγκεντρώσει απο τις προηγούμενες επαναλήψεις ή κάποιο υποσύνολο τους $\{z_i : i \in \mathcal{J}_k\}$ με $\mathcal{J}_k \subset \{1, 2, \dots, k\}$, καθώς και subgradient $w_i \in \partial f(z_i)$.
- Χρησιμοποιούμε τα δοκιμαστικά σημεία z_i για την προσέγγιση της f τοπικά απο μια γραμμική συνάρτηση της μορφής

$$\bar{f}_i(x) = f(z_i) + \langle w_i, x - z_i \rangle,$$

όπου λόγω κυρτότητας

$$f(x) \geq \bar{f}_i(x), \quad \forall x \in \mathbb{R}^d, \quad a_{i,k} := f(x_k) - \bar{f}_i(x_k) \geq 0.$$

- Για κάποιο $\epsilon_k > 0$ κατασκευάζουμε την κυρτή θήκη (convex hull)

$$C = \{w \in \mathbb{R}^d : \sum_{i \in \mathcal{J}_k} \lambda_i w_i, \lambda_i \in [0, 1], \sum_{i \in \mathcal{J}_k} \lambda_i = 1, \sum_{i \in \mathcal{J}_k} \lambda_i a_{i,k} \leq \epsilon_k\}$$

που είναι η δέσμη των subgradient για τα οποία το σφάλμα απο την 'γραμμικοποίηση' της f όπως αυτό περιγράφεται απο τα $a_{i,k}$

- Το C (όπως και το υποδιαφορικό) είναι κυρτό σύνολο και αποτελεί κατα κάποιο τρόπο προσέγγιση του υποδιαφορικού $\partial f(x_k)$, υπο την έννοια ότι

$$C \subset \partial_{\epsilon_k} f(x_k) := \{w \in \mathbb{R}^d : f(x) \geq f(x_k) + \langle w, x - x_k \rangle - \epsilon_k\}$$

με $\partial_{\epsilon_k} f(x_k) \subset \partial f(x_k)$, με τα δυο σύνολα να τείνουν το ένα στο άλλο καθώς $\epsilon_k \rightarrow 0$.

- Εφαρμόζουμε τώρα τον αλγόριθμο subgradient για την μια διεύθυνση $d_k = -w$, $w \in C$ και μαλιστα για το στοιχείο ελάχιστης νορμας του C .
- Η διεύθυνση αυτή μπορεί να καθοριστεί ως

$$d_k = \arg \min_{z \in \mathbb{R}^d} \left(\hat{f}(x_k + z) - \frac{1}{2} \alpha_k \|z\|^2 \right)$$

όπου

$$\hat{f}(x) = \max_{i \in \mathcal{J}_k} \{\bar{f}_i(x)\}$$

$\alpha_k > 0$ είναι μια κατάλληλη παράμετρος ομαλοποίησης.

- Για τον υπολογισμό του βήματος κατά την διευθυνση d_k εργαζομαστε ως εξής
- Επιλέγουμε 3 αριθμούς $m_L, m_R, \bar{t}$, που ικανοποιούν

$$m_L \in (0, \frac{1}{2}), \quad m_R \in (m_L, 1), \quad \bar{t} \in (0, 1]$$

- Επιλέγουμε το μεγαλύτερο $t_{L,k}$ τέτοιο ώστε

$$f(x_k + t_{k,L}d_k) \leq f(x_k) + m_L t_{k,L}(\hat{f}_k(x_k + d_k) - f(x_k))$$

- Αν $t_{k,L} \in [\bar{t}, 1]$ τότε

$$x_{k+1} = x_k + t_{k,L}d_k, \quad \text{και} \quad y_{k+1} = x_{k+1}$$

- Αν $t_{k,L} \in (0, \bar{t})$ τότε επιλέγουμε $t_{k,R} > t_{k,L}$ έτσι ώστε

$$-a_{k,k}$$

$$x_{k+1} = x_k + t_{k,L}d_k, \quad \text{και} \quad y_{k+1} = x_{k+1}$$