Algorytmy Statystyki Praktycznej

Szymon Majewski
smajewski@impan.pl

IMPAN

12 marca 2019

Przypomnienie

Wykład rozpoczeliśmy od omawiania metod rozwiązywania problemów optymalizacyjnych występujących w statystyce obliczeniowej. Dla wielu procedur statystycznych, wielkość którą chcemy obliczyć można wyrazić jako ekstremum pewnej funkcji:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathrm{L}(\theta).$$

Na ostatnim wykładzie omawiane były:

- Regresja liniowa i MNK
- Rozwiązywanie równań liniowych i rozkłady macierzy
- Metoda Newtona-Rhapsona
- Regresja logistyczna i GLM

Plan na dzisiaj

Wstęp do optymalizacji wypukłej. Analiza wypukła:

- Przypomnienie podstawowych informacji i definicji dotyczących zbiorów i funkcji wypukłych
- Subgradient, subróżniczka, stożki styczne i ich własności
- Warunki konieczne na ekstrema funkcji wypukłych
- Rzutowania na zbiory wypkłe i operator prox.

Algorytmy optymalizacyjne:

- Spadek gradientowy
- Metoda subgradientu
- Wersje rzutowane
- Proximal gradient

Materiały

- Optymalizacja 2, wykłady 1-5.
  http://mst.mimuw.edu.pl/lecture.php?lecture=op2
- Stephen Boyd, Lieven Vandenbherge, Convex optimization (dostępne w internecie za darmo)
- Neal Parikh, Stephen Boyd, Proximal Algorithms. Praca przeglądowa, dostępna za darmo w internecie.

Zbiory wypukłe

Zbiór W będący podzbiorem przestrzeni liniowej L nad ciałem $\mathbb{R}$ nazywamy wypukłym, jeśli dla dowolnych dwóch punktów $x, y \in W$ oraz $\lambda \in [0, 1]$ mamy:

$$\lambda x + (1 - \lambda) y \in W$$

Innymi słowy, zbiór nazywamy wypukłym wtedy i tylko wtedy gdy zawiera wszystkie odcinki o końcach należących do tego zbioru.

Funkcje wypukłe

Funkcję $f : W \to \mathbb{R}$ określoną na zbiorze wypukłym W nazwiemy wypukłą, wtedy i tylko wtedy gdy dla wszystkich $x, y \in W$ oraz $\lambda \in [0, 1]$ mamy:

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y)$$

Jeśli nierówność jest ostra dla $\lambda \in (0, 1)$, to funkcję nazywamy ściśle wypukłą.

Optymalizacja wypukła

Szukamy rozwiązania problemu:

$$x_* = \arg\min_{x \in W} f(x)$$

gdzie $f : W \to \mathbb{R}$ jest funkcja wypukłą określoną na zbiorze wypukłym W. Ograniczymy się do $W \subseteq \mathbb{R}^n$, chociaż wiele wyników łatwo uogólnić na ośrodkowe przestrzenie Hilberta. Zagadnienie maksymalizacji funkcji wypuklej na zbiorze wypukłym jest z punktu widzenia zastosowań w statystyce obliczeniowej mało ciekawe.

Znane nam problemy

Dwa problemy które pojawiły się na ostatnim wykładzie były problemami optymalizacji wypukłej. Dla regresji liniowej szukaliśmy:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^n} \mathrm{RSS}(\beta) = \arg\min_{\beta \in \mathbb{R}^n} \|Y - X\beta\|^2$$

Dla regresji logistycznej, log-wiarogodność wyrażała się wzorem:

$$\ell(\beta) = \sum_{i=1}^{n} Y_i x_i^T \beta - \sum_{i=1}^{n} \log(1 + \exp(x_i^T \beta))$$

Szukaliśmy rozwiązania następującego problemu:

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta) = \arg\min_{\beta} -\ell(\beta)$$

Słabe twierdzenie o oddzielaniu

Poniższe twierdzenia można udowodnić elementarnymi metodami, ale dowody pominiemy. Zaczniemy od tzw. słabego twierdzenia o oddzielaniu:

Theorem

Niech $U, V \subseteq \mathbb{R}^n$ będą niepustymi zbiorami wypukłymi, takimi że $\mathrm{int}(U) \neq \emptyset$ oraz $\mathrm{int}(U) \cap V = \emptyset$. Wówczas istnieje niezerowy wektor $a \in \mathbb{R}$, taki że:

$$\sup_{x \in U} a^T x \leqslant \inf_{y \in V} a^T y$$

Epigraf funkcji

Epigrafem funkcji $f : W \to \mathbb{R}$ nazywamy zbiór

$$\mathrm{epi}(f) = \{(x, z) \subseteq W \times \mathbb{R} : z \geqslant f(x)\}$$

Proste ćwiczenie:

Theorem

Jeśli W jest wypukły, to epigraf f jest wypukły wtedy i tylko wtedy gdy f jest wypukła.

Hiperpłaszczyzny podpierające

Theorem

Jeśli f jest wypukła, to w każdym punkcie $x \in \int(W)$ istnieje hiperpłaszczyzna podpierająca, to znaczy taki wektor ξ że:

$$f(y) \geqslant f(x) + \xi^T(y - x)$$

dla wszystkich $y \in W$. Jeśli f jest różniczkowalna w x, to można przyjąć $\xi = \nabla f(x)$.

Dowód

Stosujemy słabe tw. o oddzielaniu do $\mathrm{epi}(f), \{x, f(x)\}$. Istnieje niezerowy wektor (ξ, a) taki że:

$$\xi^t y + \alpha z \leqslant \xi^T x + \alpha f(x)$$

dla $(y, z) \in \mathrm{epi}(f)$.
Zatem musi zachodzić $a \leqslant 0$. Wykażemy że $a \neq 0$. W przeciwnym wypadku mielibyśmy $\xi^T(y - x) \leqslant 0$ dla wszystkich $y \in W$, co jest niemożliwe bo $x \in \mathrm{int}(W)$.
Przeskalowując tak żeby $a = -1$, oraz biorąc $z = f(y)$ otrzymujemy:

$$f(y) \geqslant f(x) + \xi^T(y - x)$$

Charakteryzacja funkcji wypukłych

Poniższe twierdzenie w połączeniu z twierdzeniem o hiperpłaszczyźnie podpierającej daje nam charakteryzację funkcji wypukłych:

Theorem

Niech $W \subseteq \mathbb{R}^n$ będzie wypukły o niepustym wnętrzu. Jeśli dla każdego $x \in \mathrm{int}(W)$ istnieje hiperpłaszczyzna podpierajaca, to f jest wypukła.

Dowód na tablicy.

Subgradienty i subróżniczki

Dla funkcji $\mathrm{f} : \mathrm{W} \to \mathbb{R}$ subgradientem f w x nazywamy dowolny wektor ξ, taki że dla wszystkich $\mathrm{y} \in \mathrm{W}$:

$$\mathrm{f}(\mathrm{y}) \geqslant \mathrm{f}(\mathrm{x}) + \xi(\mathrm{y} - \mathrm{x})$$

Subróżniczką f w x nazywamy zbiór wszystkich subgradientów f w x. Subróżniczkę f w x będziemy oznaczać $\partial \mathrm{f}(\mathrm{x})$.

Subgradienty i subróżniczki

Z udowodnionych wcześniej twierdzeń mamy następujący wniosek:

Theorem

Jeśli W jest zbiorem wypukłym o niepustym wnętrzu, to $f : W \to \mathbb{R}$ jest wypukła wtedy i tylko wtedy gdy dla każego $x \in \mathrm{int}(W)$ zachodzi:

$$\partial f(x) \neq \emptyset$$

Lemma

Dla $f : W \to \mathbb{R}$ wypukłej i dowolnegeo $x \in W$ zbiór $\partial f(x)$ jest wypukły oraz domknięty. Ponadto, jeśli $x \in \mathrm{int}(W)$ to zbiór ten jest też ograniczony.

Dowód.

Wypukłość i domkniętość są trywialne. Dla ograniczoności przez sprzeczność, patrzymy na supremum $f(y)$ po kuli otwartej o środku w x. $\square$

Subgradienty i subróżniczki

Często przydaje się również poniższy fakt:

Lemma

Jeśli $f, g : W \to \mathbb{R}$ oraz $x \in \mathrm{int}(W)$, to mamy:

$$\partial(f + g)(x) = \partial f(x) + \partial g(x)$$

oraz dla $\lambda > 0$:

$$\partial(\lambda f)(x) = \lambda \partial f(x)$$

Dowód.

Skalowanie jest trywialne. Zawieranie $\partial f(x) + \partial g(x) \subseteq \partial(f + g)(x)$ również. Zawieranie w drugą stronę wymaga silnej wersji twierdzenia o oddzielaniu, odsyłam do notatek z Optymalizacji 2. $\square$

Stożek styczny i normalny

Definicja

Dla zbioru wypukłego W stożkiem stycznym do W w $x \in W$ nazywamy zbiór:

$$T_W(x) = \mathrm{cl}\left(\left\{a\frac{y-x}{||y-x||} : a \geqslant 0, y \in W\right\}\right)$$

Definicja

Stożkiem normalnym nazywamny natomiast stożek dualny do stożka stycznego, to znaczy:

$$N_W(x) = \{y \in \mathbb{R}^n : \forall_{x \in T_W(x)} \ y^T x \leqslant 0\}$$

Uwaga: Zbiór $N_W(x)$ jest subróżniczką funkcji charakterystycznej W!

Zbiór rozwiązań

Zauważmy najpierw, że zbiór:

$$X_* = \{x : f(x) = \arg\min_{y \in W} f(y)\}$$

może być pusty lub zawierać nieskończenie wiele elementów. Zachodzi jednak zawsze:

Lemat

Zbiór X_* jest wypukły

Dowód: Jeśli $x \in X_*$ oraz $y \in X_*$, to
$f(\lambda x + (1-\lambda)y) \leqslant \arg\min_{y \in W} f(y)$, więc $\lambda x + (1-\lambda)y \in X_*$

Warunki optymalności

Udowodnimy teraz dwa podstawowe warunki konieczne optymalności.

Theorem

Dla funkcji wypukłej $f : \mathbb{R}^n \to \mathbb{R}$ mamy $x \in X_*$ wtedy i tylko wtedy gdy $0 \in \partial f(x)$.

Dowód: Z definicji $0 \in \partial f(x)$ wtedy i tylko wtedy gdy $f(y) \geqslant f(x)$ dla wszystkich $y \in \mathbb{R}^n$.

Warunki optymalności

Dla funkcji wypukłych określonych na zbiorze ograniczonym, trzeba ten warunek odpowiednio zmodyfikować:

Theorem

Niech V, W będą zbiorami wypukłymi, $V \subseteq \mathrm{int}(W)$, oraz niech $f : W \to \mathbb{R}$ będzie wypukła. Mamy $x \in \arg\min_{y \in V} f(y)$ wtedy i tylko wtedy gdy $0 \in \partial f(x) + N_V(x)$ (czyli $-N_V(x) \cap f(x) \neq \emptyset$).

Dowód w notatkach.

Rzutowanie na zbiory wypukłe

Często w algorytmach optymalizacji z ograniczeniami będziemy wykorzystywali operator rzutowania na dany zbiór wypukły W. Jeśli $W \subseteq \mathbb{R}^n$ jest wypukłym zbiorem domkniętym, to określamy:

$$P(x) = \arg\min_{y \in W} ||x - y||_2$$

punt $P(x)$ jest wyznaczony jednoznacznie.

Operator prox

Operator prox jest bardzo ważnym narzędziem w optymalizacji funkcji nieróżniczkowalnych. Można go rozumieć jako swego rodzaju uogólnienie operatora rzutowania.
Dla funkcji $\mathrm{f} : \mathrm{W} \to \mathbb{R}$ definiujemy:

$$\mathrm{prox}_{\mathrm{f}}(\mathrm{x}) = \arg\min_{\mathrm{y} \in \mathrm{W}} \mathrm{f}(\mathrm{y}) + \frac{1}{2}||\mathrm{y} - \mathrm{x}||_2^2$$

Jeśli za f przyjmiemy funkcję charakterystyczną W, to otrzymamy operator rzutowania.

Norma L1

Podstawowy przykład funkcji niegładkiej dla której operator rzutowania jest łatwy do policzenia do norma L1. Dla $\mathrm{f}(\mathrm{x}) = \lambda \sum_{\mathrm{i}=1}^{\mathrm{d}} |\mathrm{x_i}|$ mamy:

$$(\mathrm{prox_f}(\mathrm{x}))_\mathrm{i} = \mathrm{sign}(\mathrm{x_i}) \max(\mathrm{x_i} - \lambda, 0)$$

Warunek na prox

Theorem

Dla funkcji wypukłej $f : \mathbb{R}^n \to \mathbb{R}$ mamy $y = \mathrm{prox}(x)$ wtedy i tylko wtedy gdy:

$$x - y \in \partial f(y)$$

Dowód.

Można przeprowadzić dowód jak dla warunków optymalności. Można też skorzystać z faktu o subróżniczce sumy i warunków optymalności. $\square$

Algorytmy

Metody iteracyjne rozwiązują problem optymalizacyjny konstrując iteracyjnie ciąg $\{x_k\}_{k=0}^{\infty}$, tak długo aż problem optymalizacyjny zostanie rozwiązany z ustaloną dokładnością (np. algorytm Newtona - Rhapsona).
Złożoność obliczeniowa takich metod jest wypadkową dwóch czynników:

- Złożoności obliczeniowej jednej iteracji
- Ilości iteracji potrzebnych do uzyskania ϵ-dokładnego rozwiązania.

Skupimy się na omawianiu metod dla których koszt pojedynczej iteracji jest liniowy (lub niewiele większy niż liniowy) ze względu na wymiar.

Spadek gradientowy

Problem:

$$\arg\min_{x \in \mathbb{R}^n} f(x),$$

f -wypukła, klasy C^1.

Algorytm:

$$x_k = x_{k-1} - \gamma_k \nabla f(x_{k-1})$$

- Jakie długości kroków wybrać? (stałe, malejące)
- Czy i jak szybko zbiega ten algorytm?

Silna wypukłość

Definicja

Funkcję f klasy C^1 nazywamy m-silnie wypukłą, dla $m \geqslant 0$ wtedy i tylko wtedy gdy

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|x - y\|^2.$$

Uwagi:

- Równoważnie silną wypukłość można zdefiniować przez

$$f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y) - \frac{1}{2}mt(1-t)\|x - y\|^2$$

- f jest m silnie wypukła wtedy i tylko wtedy gdy $f(x) - \frac{m}{2}\|x\|^2$ jest wypukła.
- Gdy $m = 0$ silna wypukłość oznacza zwyczajną wypukłość.

L-gładkość

Definicja

Funkcje f nazywamy L-gładką jeżeli jej gradient ∇f jest Lipshitzowski ze stałą L:

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

Stwierdzenie

Jeżeli f jest L-gładka to dla dowolnego x, y zachodzi

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{1}$$

Dowód

Zdefiniujmy $g(t) = f(x + t(y - x))$. Wówczas $g'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$ oraz

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt.$$

Stąd

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle - \langle \nabla f(x), y - x \rangle dt \\ &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ &\leqslant \|y - x\| \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| dt \\ &\leqslant L\|y - x\|^2 \int_0^1 t dt = \frac{L}{2}\|y - x\|^2 \end{aligned}$$

Monotoniczność spadku gradientowego

$$x_k = x_{k-1} - \gamma_k \nabla f(x_{k-1})$$

Lemat (o monotoniczności)

Jeśli nierówność (1) jest spełniona dla $y = x_k$ $x = x_{k-1}$ oraz $L = \frac{1}{\gamma_k}$ to

$$f(x_k) \leqslant f(x_{k-1})$$

Dowód

Zdefiniujmy $Q_\gamma(x, y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\gamma}\|y - x\|^2$.

- $Q_\gamma(x, x) = f(x)$
- $\arg\min_y Q_{\gamma_k}(x_{k-1}, y) = x_{k-1} - \gamma_k \nabla f(x_{k-1}) = x_k$
- Jeśli $\gamma \leqslant L^{-1}$ to

$$Q_\gamma(x, y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \geqslant f(y)$$

Czyli

$$f(x_{k-1}) = Q(x_{k-1}, x_{k-1}) \geqslant Q(x_{k-1}, x_k) \geqslant f(x_k).$$

Zbieżność spadku po gradiencie

TWIERDZENIE

Jeśli f jest wypukła , L-gładka i osiąga minimum w x^* oraz dla pewnego $a > 0$ zachodzi $aL^{-1} \leqslant \gamma_k$ oraz spełnione są założenia Lematu (2) to dla dowolnego k

$$f(x_k) - f(x^*) \leqslant \frac{L\|x_0 - x^*\|^2}{2ak}.$$

Ponadto jeśli f jest m-silnie wypukła to

$$\|x_k - x^*\|^2 \leqslant \left(1 - \frac{ma}{L}\right)^k \|x_0 - x^*\|^2.$$

Dowód I

Lemat

Jeśli f jest m-silnie wypukła, spełnione są założenia lematu o monotoniczności, oraz dla pewnego $a > 0$ zachodzi $aL^{-1} \leqslant \gamma_k$ to dla dowolnego k i dowolnego y

$$2\gamma_k(f(x_k) - f(y)) \leqslant \left(1 - \frac{ma}{L}\right) \|x_{k-1} - y\|^2 - \|x_k - y\|^2$$

Dowód twierdzenia

Niech x^* punkt w którym f osiąga minimum wtedy $f(x_k) - f(x^*) > 0$ i z lematu otrzymujemy

$$0 \leqslant \left(1 - \frac{ma}{L}\right) \|x_{k-1} - y\|^2 - \|x_k - y\|^2$$

Czyli

$$\|x_k - x^*\|^2 \leqslant \left(1 - \frac{ma}{L}\right) \|x_{k-1} - x^*\|^2$$

Dowód II

Iterując tą nierównośc otrzymujemy 2 część twierdzenia.

Z lematu otrzymujemy

$$2\gamma_i(f(x_i) - f(x^*)) \leqslant \|x_{i-1} - x^*\|^2 - \|x_i - x^*\|^2$$

Sumujemy nierówności od $i = 1$ do k

$$\sum_{i=1}^{k} 2\gamma_i(f(x_i) - f(x^*)) \leqslant \|x_0 - x^*\|^2 - \|x_k - x^*\|^2$$

Z monotoniczności $f(x_i)$ oraz z $\gamma_i \geqslant aL^{-1}$ otrzymujemy

$$\frac{2ak}{L}(f(x_k) - f(x^*)) \leqslant \|x_0 - x^*\|^2 - \|x_k - x^*\|^2$$

Dowód lematu I

Z nierówności (1) mamy dla $L = \frac{1}{\gamma_k}$

$$
\begin{aligned}
f(x_k) &\leqslant f(x_{k-1}) + \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle + \frac{L}{2}\|x_k - x_{k-1}\|^2 \\
&= f(x_{k-1}) - \frac{1}{\gamma_k}\|x_k - x_{k+1}\|^2 + \frac{L}{2}\|x_k - x_{k-1}\|^2 \\
&= f(x_{k-1}) - \frac{1}{\gamma_k}\left(1 - \frac{L\gamma_k}{2}\right)\|x_k - x_{k-1}\|^2 \\
&= f(x_{k-1}) - \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2
\end{aligned}
$$

Z wypukłości

$$
\begin{aligned}
f(y) &\geqslant f(x_{k-1}) + \langle \nabla f(x_k), y - x_{k-1} \rangle + \frac{m}{2}\|x_{k-1} - y\|^2 \\
&= f(x_{k-1}) - \frac{1}{\gamma_k}\langle x_k - x_{k-1}, y - x_{k-1} \rangle + \frac{m}{2}\|x_{k-1} - y\|^2
\end{aligned}
$$

Dowód lematu II

Mnożąc nierówności przez $2\gamma_k$ i odejmując stronami otrzymujemy.

$$\begin{aligned}
2\gamma_k(f(x_k) - f(y)) &\leqslant 2\langle x_k - x_{k-1}, y - x_{k-1}\rangle - \|x_k - x_{k-1}\|^2 - \gamma_k m\|x_{k-1} - y\|^2 \\
&= (1 - m\gamma_k)\|x_{k-1} - y\|^2 - \|x_{k-1} - y\|^2 + 2\langle x_k - x_{k-1}, y - x_{k-1}\rangle - \|x_k - x_{k-1}\|^2 \\
&= (1 - m\gamma_k)\|x_{k-1} - y\|^2 - \|x_{k-1} - y - (x_{k-1} - x_k)\|^2 \\
&= (1 - m\gamma_k)\|x_{k-1} - y\|^2 - \|x_k - y\|^2
\end{aligned}$$

Jak wybrać γ_k

- Optymalne jest wybrać $\gamma_k = L$, jednak obliczenie L (dobre oszacowanie) może, być truden lub niemożliwe.
- Z dowodu wynika, że jeśli $\gamma_k \to 0$ oraz $\sum_{k=0}^{\infty} \gamma_k = \infty$ to tez otrzymamy zbieżność ale w tempie zależnym od tempa zbieżności γ_k.
- Innym rozwiązaniem jest backtracking. Ustalamy γ_0 i wybieramy $0 < \eta < 1$ I wykonujemy algorytm
  1. $\gamma_k = \gamma_{k-1}$
  2. $x_k = x_{k-1} - \gamma_k \nabla f(x_{k-1})$
  3. Jeśli

     $$f(x_k) \leqslant f(x_{k-1}) + \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle + \frac{1}{2\gamma_k} \|x_k - x_{k-1}\|^2$$

     to $k \to k+1$. W przeciwnym przypadku $\gamma_k = \eta \gamma_k$ i wracamy do kroku 2.

Rzutowany spadek gradientowy

Modyfikacja spadku gradientowego do optymalizacji na podzbiorach wypukłych:

$$x_k = P_W(x_{k-1} - \gamma_k \nabla f(x_{k-1}))$$

Analiza analogiczna do normalnego spadku gradientowego: dowodzimy lematu głównego a następnie wnioskujemy tempo zbieżności w taki sam sposób jak dla normalnego spadku gradientowego.

Rzutowany spadek gradientowy

Naśladując dowód lematu dla spadku gradientowego, otrzymujemy dla dowolnych $x_{k-1}, y \in W$:

$$2\gamma_k(f(x_k) - f(y)) \leqslant \left(1 - \frac{ma}{L}\right) \|x_{k-1} - y\|^2 - \|z_k - y\|^2$$

gdzie $z_k = x_{k-1} - \gamma_k \nabla f(x_{k-1})$ oraz $x_k = P_W(z_k)$. Następnie używamy faktu, że dla dowolnego $y \in W$ zachodzi:

$$||P_W(x) - y||_2 \leqslant ||x - y||_2$$

Metoda subgradientowa

Metoda polega na zastąpieniu w spadku subgradientowym gradientu subgradientem.

- Prosta metoda którą można stosować do ogólniejszej klasy problemów.
- Nie ma własności spadku
- Wolniejsze tempo zbieżności

Metoda subgradientowa

$$x_k = x_{k-1} - \gamma_k g_{k-1}, \quad g_{k-1} \in \partial f(x_{k-1})$$

Analiza metody subgradientowej

Podobnie jak dla spadku gradientowego, mamy fundamentalny lemat:

Lemat

Dla dowolnego $k > 0$ oraz dowolnych $x_{k-1}, y \in W$ mamy:

$$2\gamma_k(f(x_{k-1}) - f(y)) \leqslant ||x_{k-1} - y||_2^2 - ||x_k - y||_2^2 + \gamma_k^2 ||g_{k-1}||_2^2$$

Dowód.

$$\begin{aligned} ||x_{k-1} - y||_2^2 - ||x_k - y||_2^2 + \gamma_k^2 ||g_{k-1}||_2^2 &= 2\gamma_k g_{k-1}^T (x_{k-1} - y) \\ &\geqslant 2\gamma_k (f(x_{k-1}) - f(y)) \end{aligned}$$

$\square$

Analiza metody subgradientowej

Z powyższego lematu, łatwo wywnioskować następujące własności. Załóżmy że f jest G-Lipshitzowska, oraz $\|x_o - x^*\| \leqslant R$. Oznaczamy $x_k^N = \left(\sum_{i=0}^{N-1} \gamma_{i+1} x_i\right) \left(\sum_{i=1}^{N} \gamma_i\right)$. Wówczas:

- Dla $\gamma_k = \gamma$
$$\lim_{k\to\infty} f(x_k^N) \leqslant f(x^*) + G^2\gamma$$
- Jeśli $\gamma_k \to 0, \sum \gamma_k^2 < \infty$ oraz $\sum \gamma_k = \infty$ to
$$\lim_{k\to\infty} f(x_k^N) = f(x^*)$$
- Jeśli $\gamma_k = \frac{R}{G\sqrt{k}}$
$$f(x_k^N) - f(x^*) \leqslant \frac{RG}{\sqrt{k}}$$

Rzutowana metoda subgradientowa

Dowodzimy lematu dla wersji z rzutowaniem. Jeśli oznaczymy $z_k = x_{k-1} - \gamma_k g_{k-1}$ oraz $x_k = P(z_k)$, to mamy:

$$2\gamma_k(f(x_{k-1} - f(y)) \leqslant ||x_{k-1} - y||_2^2 - ||z_k - y||_2^2 + \gamma_k^2 ||g_{k-1}||_2^2$$

i korzystamy z faktu że $||P(x) - y||_2 \leqslant ||x - y||_2$ dla dowolnego x oraz $y \in W$. Dalej analiza analogicznie.

LASSO

Model liniowy:

$$\mathrm{Y} = \mathrm{X}\beta + \epsilon$$

Metoda najmniejszych kwadratów:

$$\arg\min_{\beta} \|\mathrm{Y} - \mathrm{X}\beta\|^2$$

LASSO:

$$\arg\min_{\beta} \{\|\mathrm{Y} - \mathrm{X}\beta\|^2 + \lambda \sum_{\mathrm{i}=1}^{\mathrm{p}} |\beta_{\mathrm{i}}|\}$$

Jak obliczyć rozwiązanie LASSO?

Proximal gradient

Problem:

$$\arg\min_{x}\{f(x) + g(x)\}$$

- f wypukła i gładka
- g wypukła

Algorytm

$$x_{k+1} = \mathrm{prox}_{\gamma_k, g}(x_k - \gamma_k \nabla f(x_k))$$

Stąd

$$x_{k+1} \in x_k - \gamma_k \nabla f(x_k) - \gamma_k \partial g(x_{k+1})$$

Własności proximal gradient

Jeśli f jest L-gładka oraz $\gamma \leqslant L^{-1}$ to

$$f(x)+g(x) \leqslant f(y)+\nabla f(y)^T(x-y)+\frac{1}{2\gamma}\|y-x\|^2+g(x) = Q_\gamma(x,y)$$

Przekształacając

$$Q_\gamma(x,y) = f(y) - \frac{\gamma}{2}\|\nabla f(y)\|^2 + g(x) + \frac{1}{2\gamma}\|x-(y-\gamma\nabla f(y))\|^2$$

Czyli

$$\mathrm{prox}_{\gamma_k,g}(x_k - \gamma_k \nabla f(x_k)) = \arg\min_x Q_{\gamma_k}(x, x_k)$$

Zatem

$$\begin{aligned} f(x_k) + g(x_k) = Q_{\gamma_k}(x_k, x_k) &\geqslant \arg\min_x Q_{\gamma_k}(x, x_k) \\ &= Q_{\gamma_k}(x_{k+1}, x_k) \geqslant f(x_{k+1}) + g(x_{k+1}) \end{aligned}$$

Zbieżność proximal gradient

Oznaczmy przez $F = f + g$

Lemat

Jeśli f jest wypukła i L-gładka, g jest wypukła oraz $\gamma_k \leqslant L^{-1}$ to dla dowolnego y

$$2\gamma_k(F(x_k) - F(y)) \leqslant \|x_{k-1} - y\|^2 - \|x_k - y\|^2$$

Dowód:

Niech $v \in \partial g(x_k)$ takie, że

$$x_k = x_{k-1} - \gamma_k \nabla f(x_{k-1}) - \gamma_k v$$

Mamy trzy nierówności

$$F(x_k) \leqslant g(x_k) + f(x_{k-1}) + \nabla f(x_{k-1})^T (x_k - x_{k-1}) + \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2$$

$$f(y) \geqslant f(x_{k-1}) + \nabla f(x_{k-1})^T (y - x_{k-1})$$

$$g(y) \geqslant g(x_k) + v^T (y - x_k)$$

Dowód lematu c.d.

Odejmując nierówności otrzymujemy

$$F(x_k) - F(y) \leqslant \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2 + \nabla f(x_{k-1})^T(x_k - x_{k-1}) \\ - \nabla f(x_{k-1})^T(y - x_{k-1}) - v^T(y - x_k)$$

Przekształcając prawą stronę otrzymujemy

$$F(x_k) - F(y) \leqslant \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2 + (\nabla f(x_{k-1}) + v)^T(x_k - y)$$

$$F(x_k) - F(y) \leqslant \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2 - \frac{1}{\gamma_k}(x_k - x_{k-1})^T(x_k - y)$$

$$2\gamma_k(F(x_k) - F(y)) \leqslant \|x_k - x_{k-1} - x_k + y\|^2 - \|x_k - y\|^2$$

Zbieżność proximal gradient

TWIERDZENIE

Niech $F = f + g$, gdzi f jest wypukła i L-gładka, g jest wypukła oraz F osiąga minimum w x^*. Jeśli $\gamma_k = \gamma < L^{-1}$ to

$$F(x_k) - F(x^*) \leqslant \frac{L\|x_0 - x^*\|^2}{2k}.$$

Materiały

- Optymalizacja 2, wykłady 1-5.
  http://mst.mimuw.edu.pl/lecture.php?lecture=op2
- Stephen Boyd, Lieven Vandenbherge, Convex optimization (dostępne w internecie za darmo)
- Neal Parikh, Stephen Boyd, Proximal Algorithms. Praca przeglądowa, dostępna za darmo w internecie.