Algorytmy Statystyki Praktycznej

Błażej Miasojedow

MIMUW

13 marca 2019

Przypomnienie

Na poprzednim wykładzie omawiane były

- Zbiory i funkcje wypukłe.
- Subgradient i subróżniczka.
- Problemy optymalizacji wypukłej.
- Rzutowanie na zbiory wypukłe oraz operator prox
- Spadek po gradiencie

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

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

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

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

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 (1) 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.

Przyspieszenie Nesterova

Można pokazać, że optymalne tempo zbieżności algorytmów pierwszego rzędu (opartych na gradiencie) jest $\mathcal{O}(1/k^2)$. Spadek po gradiencie zbiega w tempie $\mathcal{O}(1/k)$. Czy można go poprawić?

Nesterov 1983

Niech f wypukła

$$\arg\min_x f(x)$$

1. $y_0 = x_0, t_0 = 0$
2. $x_k = y_{k-1} - \gamma_k \nabla f(y_{k-1})$
3. $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$
4. $y_k = x_k + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1})$

Przyspieszenie Nesterowa

TWIERDZENIE

Niech f wypukła, L-gładka oraz osiąga minimum w x^*. Jeśli $\gamma_k = \gamma \leqslant L^{-1}$ to

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

Dowód I

Tak jak w poprzednim dowodzie z wypukłości i L-gładkości

$$f(x_k) \leqslant f(y_{k-1}) - \frac{1}{2\gamma}\|x_k - y_{k-1}\|^2$$
$$f(z) \geqslant f(y_{k-1}) - \frac{1}{\gamma}\langle x_k - y_{k-1}, z - y_{k-1}\rangle$$

Odejmując stronami uzyskujemy

$$f(x_k) - f(z) \leqslant -\frac{1}{2\gamma}\|x_k - y_{k-1}\|^2 + \frac{1}{\gamma}\langle x_k - y_{k-1}, z - y_{k-1}\rangle$$

Uzywamy tej nierówności z $z = x_{k-1}$ i $z = x^*$

$$f(x_k) - f(x_{k-1}) \leqslant -\frac{1}{2\gamma}\|x_k - y_{k-1}\|^2 + \frac{1}{\gamma}\langle x_k - y_{k-1}, x_{k-1} - y_{k-1}\rangle$$
$$f(x_k) - f(x^*) \leqslant -\frac{1}{2\gamma}\|x_k - y_{k-1}\|^2 + \frac{1}{\gamma}\langle x_k - y_{k-1}, x^* - y_{k-1}\rangle$$

Dowód II

Mnożymy pierwszą nierównośc przez $t_{k-1} - 1$ i dodajemy do drugiej

$$t_{k-1}(f(x_k) - f(x^*)) - (t_{k-1} - 1)(f(x_{k-1}) - f(x^*)) \leqslant$$
$$-\frac{t_{k-1}}{2\gamma}\|x_k - y_{k-1}\|^2 + \frac{1}{\gamma}\langle x_k - y_{k-1}, (t_{k-1} - 1)x_{k-1} - t_{k-1}y_{k-1} + x^*\rangle$$

Oznaczmy przez $\Delta_k = f(x_k) - f(x^*)$ i pomnóżmy nierówmość przez t_{k-1} $(t_{k-2}^2 = t_{k-1}^2 - t_{k-1})$

$$t_{k-1}^2\Delta_k - t_{k-2}^2\Delta_{k-1} \leqslant \frac{-1}{2\gamma}\Big[\|(x_k - y_{k-1})t_{k-1}\|^2$$
$$+ 2t_{k-1}\langle x_k - y_{k-1}, t_{k-1}y_{k-1} - (t_{k-1} - 1)x_{k-1} - x^*\rangle\Big]$$

Dowód III

Po przekształceniu

$$t_{k-1}^2\Delta_k - t_{k-2}^2\Delta_{k-1} \leqslant \frac{-1}{2\gamma}$$

$$\left[\|x_k t_{k-1} - (t_{k-1}-1)x_{k-1} - x^*\|^2 - \|t_{k-1}y_{k-1} - (t_{k-1}-1)x_{k-1} - x^*\|^2\right]$$

Dodatkowo

$$\begin{aligned} y_k &= x_k + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1}) \\ t_k y_k &= t_k x_k + (t_{k-1}-1)(x_k - x_{k-1}) \\ t_k y_k - (t_k - 1)x_k &= t_{k-1}x_k - (t_{k-1}-1)x_{k-1} \end{aligned}$$

Oznaczając przez $u_k = t_{k-1}x_k - (t_{k-1}-1)x_{k-1} - x^*$
Stąd

$$t_{k-1}^2\Delta_k - t_{k-2}^2\Delta_{k-1} \leqslant \frac{1}{2\gamma}\left(\|u_{k-1}\|^2 - \|u_k\|^2\right)$$

Dowód IV

Sumując nierówność od $i = 2$ do k otrzymujemy

$$t_{k-1}^2 \Delta_k \leqslant \frac{1}{2\gamma}\left(\|u_1\|^2 - \|u_k\|^2\right) \leqslant \frac{1}{2\gamma}\|u_1\|^2$$

Gdzie

$$u_1 = t_0 x_k - (t_0 - 1)x_0 - x^* = x_0 - x^*$$

Oraz dla $k \geqslant 2$ (Indukcja)

$$t_{k-1} \geqslant \frac{k}{2}$$

Backtracking

1. $y_0 = x_0, t_0 = 0$, γ_0 i $0 < \eta < 1$
2. $\gamma_k = \gamma_{k-1}$
3. $x_k = y_{k-1} - \gamma_k \nabla f(y_{k-1})$
4. Jeśli
$$f(x_k) \leqslant f(y_{k-1}) + \langle \nabla f(y_{k-1}), x_k - y_{k-1} \rangle + \frac{1}{2\gamma_k} \|x_k - y_{k-1}\|^2$$
to krok 5. W przeciwnym przypadku $\gamma_k = \eta \gamma_k$ i wracamy do kroku 3.
5. $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$
6. $y_k = x_k + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1})$

TWIERDZENIE

Niech f wypukła, L-gładka oraz osiąga minimum w x^*. Jeśli γ_k jak powyżej to

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

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

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

Przyspieszenie Nesterova dla proximal gradient

1. $y_0 = x_0, t_0 = 0$, γ_0 i $0 < \eta < 1$
2. $\gamma_k = \gamma_{k-1}$
3. $x_k = \mathrm{prox}_{\gamma_k g}(y_{k-1} - \gamma_k \nabla f(y_{k-1}))$
4. Jeśli
$$f(x_k) \leqslant f(y_{k-1}) + \langle \nabla f(y_{k-1}), x_k - y_{k-1} \rangle + \frac{1}{2\gamma_k} \|x_k - y_{k-1}\|^2$$
to krok 5. W przeciwnym przypadku $\gamma_k = \eta \gamma_k$ i wracamy do kroku 3.
5. $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$
6. $y_k = x_k + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1})$

TWIERDZENIE

Niech f wypukła, L-gładka, g wypukła oraz F osiąga minimum w x^*. Jeśli γ_k jak powyżej to

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

Alternating Direction Method of Multipliers

Rozważmy problem

$$\min_{x:\ Ax=b} f(x)$$

Lagrangian dla tego problemu

$$\mathcal{L}(x, \mu) = f(x) + \mu^T(Ax - b)$$

Problem dualny

$$\max_{\mu} g(\mu), \quad g(\mu) = \inf_{x} \mathcal{L}(x, \mu)$$

Rozwiązanie

$$x^* = \arg\min_{x} \mathcal{L}(x, \mu^*)$$

Algorytm gradientowy dla problemu dualnego

$$\mu_{k+1} = \mu_k + \gamma_k \nabla g(\mu_k),$$

gdzie

$$\nabla g(\mu_k) = A\tilde{x} - b, \quad \tilde{x} = \arg\min \mathcal{L}(x, \mu_k).$$

Czyli

$$\begin{aligned} x_{k+1} &= \arg\min \mathcal{L}(x, \mu_k) \\ \mu_{k+1} &= \mu_k + \gamma_k (Ax_{k+1} - b) \end{aligned}$$

Algorytm jest zbieżny przy restrykcyjnych założeniach. Nie praktyczny

Metoda mnożników I

Warunki konieczne

$$\mathrm{Ax}^* - \mathrm{b} = 0, \quad \nabla\mathcal{L}(\mathrm{x}^*, \mu^*) = \nabla\mathrm{f}(\mathrm{x}^*) + \mathrm{A}^{\mathrm{T}}\mu^* = 0$$

Chcielibyśmy aby

$$\nabla\mathcal{L}(\mathrm{x}_\mathrm{k}, \mu_\mathrm{k}) = 0$$

Równoważnie, warunki konieczne można zapisać w postaci

$$\mathrm{Ax}^* - \mathrm{b} = 0, \quad \nabla\mathrm{f}(\mathrm{x}^*) + \mathrm{A}^{\mathrm{T}}\mu^* + \rho\mathrm{A}^{\mathrm{T}}(\mathrm{Ax}^* - \mathrm{b}) = 0$$

Co odpowiada

$$\mathrm{Ax}^* - \mathrm{b} = 0, \quad \nabla\mathcal{L}_\rho(\mathrm{x}^*, \mu^*) = 0,$$

gdzie

$$\mathcal{L}_\rho(\mathrm{x}, \mu) = \mathrm{f}(\mathrm{x}) + \mu^{\mathrm{T}}(\mathrm{Ax} - \mathrm{b}) + \frac{\rho}{2}\|\mathrm{Ax} - \mathrm{b}\|^2$$

Metoda mnożników II

Algorytm:

$$\begin{aligned} x_{k+1} &= \arg\min_x \mathcal{L}_\rho(x, \mu_k) \\ \mu_{k+1} &= \mu_k + \rho(Ax_{k+1} - b) \end{aligned}$$

Mamy

$$\begin{aligned} 0 &= \nabla \mathcal{L}_\rho(x_{k+1}, \mu_k) \\ &= \nabla f(x_{k+1}) + A^T \mu_k + \rho A^T (Ax_{k+1} - b) \\ &= \nabla f(x_{k+1}) + A^T (\mu_k + \rho(Ax_{k+1} - b)) \\ &= \nabla f(x_{k+1}) + A^T \mu_{k+1} \end{aligned}$$

II warunek osiągany jest asymptotycznie

$$Ax_k - b \to 0$$

ADMM

Rozważmy problem

$$\min_{\{x,z:\ Ax+Bz=c\}} f(x) + g(z)$$

f,g wypukłe

$$\mathcal{L}_\rho(x, z, \mu) = f(x) + g(z) + \mu^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2$$

Algorytm

1. $x_{k+1} = \arg\min_x \mathcal{L}_\rho(x, z_k, \mu_k)$
2. $z_{k+1} = \arg\min_z \mathcal{L}_\rho(x_{k+1}, z, \mu_k)$
3. $\mu_{k+1} = \mu_k + \rho(Ax_{k+1} + Bz_{k+1} - c)$

Uwagi

- Kroki 1 i 2 można zastąpić przez rozwiązanie numeryczne.
- Algorytm jest zbieżny jeżeli $\mathcal{L}_0$ ma punkt przegięcia.

ADMM dla LASSO

Problem

$$\min_{\beta} \frac{1}{2}\|\mathrm{Y} - \mathrm{X}\beta\|^2 + \lambda\|\beta\|_1$$

można zapisać jako

$$\min_{\{\beta,\mathrm{z}:\ \beta=\mathrm{z}\}} \frac{1}{2}\|\mathrm{Y} - \mathrm{X}\beta\|^2 + \lambda\|\mathrm{z}\|_1$$

oraz

$$\mathcal{L}_\rho(\beta, \mathrm{z}, \mu) = \frac{1}{2}\|\mathrm{Y} - \mathrm{X}\beta\|^2 + \lambda\|\mathrm{z}\|_1 + \mu^{\mathrm{T}}(\beta - \mathrm{z}) + \frac{\rho}{2}\|\beta - \mathrm{z}\|^2$$

ADMM

1. $\beta_{\mathrm{k}+1} = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \rho\mathrm{I})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{Y} - \mu_{\mathrm{k}} + \rho\mathrm{z}_{\mathrm{k}})$
2. $\mathrm{z}_{\mathrm{k}+1} = \mathrm{prox}_{\frac{\lambda}{\rho}\|\mathrm{z}\|_1}(\beta_{\mathrm{k}+1} + \frac{1}{\rho}\mu_{\mathrm{k}}) =$
   $\mathrm{sign}(\beta_{\mathrm{k}+1} + \frac{1}{\rho}\mu_{\mathrm{k}})(|\beta_{\mathrm{k}+1} + \frac{1}{\rho}\mu_{\mathrm{k}}| - \frac{\lambda}{\rho})_+$
3. $\mu_{\mathrm{k}+1} = \mu_{\mathrm{k}} + \rho(\beta_{\mathrm{k}+1} - \mathrm{z}_{\mathrm{k}+1})$

ADMM dla optymalizacji z ograniczeniami wypukłymi I

Problem:

$$\min_{x \in \mathcal{K}} f(x)$$

f- wypukła, $\mathcal{K}$ domknięty wypukły podzbiór $\mathbb{R}^d$.

$$\min_{x=z} f(x) + 1_{\mathcal{K}}(z)$$

gdzie

$$1_{\mathcal{K}}(z) = \begin{cases} 1 & z \in \mathcal{K} \\ +\infty & z \notin \mathcal{K} \end{cases}$$

$$\mathcal{L}_\rho(x, z, \mu) = f(x) + 1_{\mathcal{K}}(z) + \mu^T(x - z) + \frac{\rho}{2}\|x - z\|^2$$

ADMM

1. $x_{k+1} = \arg\min_x \mathcal{L}_\rho(x, z_k, \mu_k)$
2. $z_{k+1} = \Pi_{\mathcal{K}}(x_{k+1} + \frac{1}{\rho}\mu_k)$
3. $\mu_{k+1} = \mu_k + \rho(x_{k+1} - z_{k+1})$

ADMM dla optymalizacji z ograniczeniami wypukłymi II

Jeżeli f funkcja kwadratowa czyli

$$f(x) = \frac{1}{2}x^T P x + q^T x + r$$

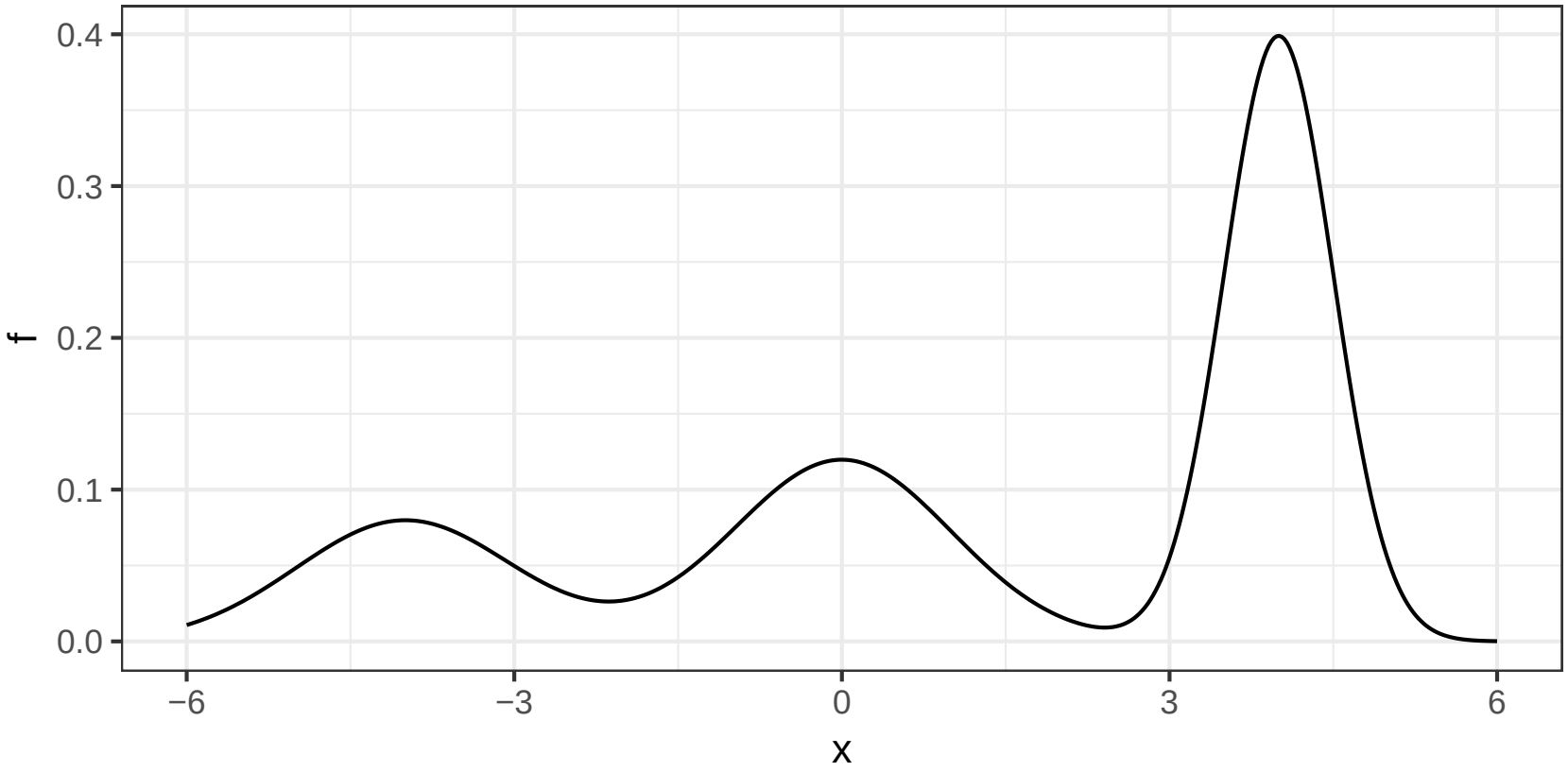
to

$$\arg\min_x \mathcal{L}_\rho(x, z_k, \mu_k) = (P + \rho I)^{-1}(\rho z_k - q - \mu_k)$$

Mieszanka rozkładów gaussowskich

Niech $Y_1, \ldots, Y_n$ będą zmiennymi losowymi o rozkładzie

$$f_\theta(y) = \sum_{i=1}^{k} \alpha_i \phi_{\mu_i,\sigma_i}(y), \quad \theta = (\alpha_i, \mu_i, \sigma_i)_{i=1}^{k}$$



Jak obliczyć estymator θ?

Mieszanka jako model z brakującymi danymi

Rozważmy parę (X, Y) o rozkładzie

$$\mathrm{P}(\mathrm{X} = \mathrm{i}) = \alpha_{\mathrm{i}}, \quad \mathrm{Y}|\mathrm{X} = \mathrm{i} \sim \mathcal{N}(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}^2)$$

Wówczas rozkład brzegowy Y ma gęstość

$$\mathrm{f}_\theta(\mathrm{y}) = \sum_{\mathrm{i}=1}^{\mathrm{k}} \alpha_{\mathrm{i}} \phi_{\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}}(\mathrm{y})$$

Expectation Maximization I

Rozkład łączny

$$\mathrm{f}_\theta(\mathrm{x}, \mathrm{y})$$

Obserwujemy Y z rozkładu brzegowego

$$\mathrm{f}_\theta(\mathrm{y}) = \int \mathrm{f}_\theta(\mathrm{x}, \mathrm{y})\mathrm{dx}$$

Chcemy policzyć MLE

$$\arg\max_\theta \ell(\theta), \quad \ell(\theta) = \log(\mathrm{f}(\theta)(\mathrm{y}))$$

Wprowadźmy funkcję pomocniczą

$$\mathrm{Q}(\theta, \vartheta) = \int \log(\mathrm{f}_\theta(\mathrm{x}, \mathrm{y}))\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y})\mathrm{dx} = \mathbb{E}_\vartheta(\log(\mathrm{f}_\theta(\mathrm{X}, \mathrm{Y}))|\mathrm{Y})$$

Expectation Maximization II

Krok Expectation:

$$\mathrm{Q}(\theta, \theta_{\mathrm{k}}) = \mathbb{E}_{\theta_{\mathrm{k}}}(\log(\mathrm{f}_{\theta}(\mathrm{X}, \mathrm{Y}))|\mathrm{Y})$$

Krok Maximization:

$$\theta_{\mathrm{k}+1} = \arg\max_{\theta} \mathrm{Q}(\theta, \theta_{\mathrm{k}})$$

Własności funkcji Q

Oznaczmy przez

$$\mathrm{H}(\theta, \vartheta) = -\int \log(\mathrm{f}_\theta(\mathrm{x}|\mathrm{y}))\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y})\mathrm{dx}$$

1. $\mathrm{Q}(\theta, \vartheta) = \ell(\theta) - \mathrm{H}(\theta, \vartheta)$
2. $\ell(\theta) - \ell(\vartheta) \geqslant \mathrm{Q}(\theta, \vartheta) - \mathrm{Q}(\vartheta, \vartheta)$
3. $\nabla_\theta \mathrm{Q}(\theta, \vartheta)|_{\theta=\vartheta} = \nabla \ell(\vartheta)$

Powyższe własności wynikają z faktu, że

$$\begin{aligned}\mathrm{H}(\theta, \vartheta) - \mathrm{H}(\vartheta, \vartheta) &= -\int \log(\mathrm{f}_\theta(\mathrm{x}|\mathrm{y}))\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y}) + -\int \log(\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y}))\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y}) \\ &= -\int \log\left(\frac{\mathrm{f}_\theta(\mathrm{x}|\mathrm{y})}{\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y})}\right)\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y})\mathrm{dx} \geqslant 0\end{aligned}$$

EM dla rodziny wykladniczej

Niech

$$\mathrm{f}_\theta(\mathrm{x},\mathrm{y}) = \exp\{\theta^{\mathrm{T}}\mathrm{S}(\mathrm{x},\mathrm{y}) - \mathrm{b}(\theta)\}$$

Zatem

$$\mathrm{Q}(\theta,\vartheta) = \mathbb{E}_\vartheta(\log(\mathrm{f}_\theta(\mathrm{X},\mathrm{Y}))|\mathrm{Y}) = \theta^{\mathrm{T}}\mathbb{E}_\vartheta(\mathrm{S}(\mathrm{X},\mathrm{Y})|\mathrm{Y}) - \mathrm{b}(\theta)$$

EM dla mieszanek I

$Y_1, \ldots, Y_n$ z rozkładu

$$f_\theta(y) = \sum_{i=1}^{k} \alpha_k \phi_{\mu_i,\sigma_i}(y)$$

Łączny rozkład $P(X_j = i) = \alpha_i$, $f_\theta(y_j|x_j = i) = \phi_{\mu_i,\sigma_i}(y_j)$

$$f_\theta(y, x) = \prod_{j=1}^{n} \prod_{i=1}^{k} [\alpha_i \phi_{\mu_i,\sigma_i}(y_j)]^{1(x_j=i)}$$

Zatem

$$Q(\theta, \theta') = \sum_{j=1}^{n} \sum_{i=1}^{k} \mathbb{E}_{\theta'}(1(X_j = i)|Y_j = y_j)[\log(\alpha_i) + \log(\phi_{\mu_i,\sigma_i}(y_j))]$$

EM dla mieszanek II

Oznaczmy przez
$w_{ij} = \mathbb{E}_{\theta'}(1(X_j = i)|Y_j = y_j) = \mathbb{P}_{\theta'}(X_j = i|Y_j = y_j)$ Ze wzoru Bayesa mamy

$$w_{ij} = \frac{\alpha'_i \phi_{\mu'_i,\sigma'_i}(y_j)}{\sum_{i=1}^{k} \alpha'_i \phi_{\mu'_i,\sigma'_i}(y_j)}$$

Stąd

$$Q(\theta, \theta') = \sum_{j=1}^{n} \sum_{i=1}^{k} w_{ij}[\log(\alpha_i) + \log(\phi_{\mu_i,\sigma_i}(y_j))]$$

Maksymalizacja po α:
Chcemy zmaksymalizować $Q(\theta, \theta')$ względem α przy ograniczeniu $\sum \alpha_i = 1$. Lagrangian dla tego problemu to

$$\mathcal{L}(\alpha, \lambda) = \sum_{j=1}^{n} \sum_{i=1}^{k} w_{ij} \log(\alpha_j) - \lambda(\sum_{i=1}^{k} \alpha_j - 1)$$

EM dla mieszanek III

Stąd

$$\frac{\partial \mathcal{L}(\alpha, \lambda)}{\partial \alpha_j} = \frac{\sum_{j=1}^{n} w_{ij}}{\alpha_j} - \lambda = 0$$

Czyli

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

Maksymalizacja po μ i σ^2:

$$g(\mu, \sigma^2) = \sum_{j=1}^{n} \sum_{i=1}^{k} w_{ij} \left[\frac{-(y_j - \mu_i)^2}{2\sigma_i^2} - \frac{1}{2} \log(\sigma_i^2) \right]$$

$$\frac{\partial g}{\partial \mu_i} = \sum_{j=1}^{n} w_{ij} \frac{y_j - \mu_i}{\sigma_i^2} = 0$$

To

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}}$$

EM dla mieszanek IV

oraz $(\tau_i = \sigma_i^2)$

$$\frac{\partial g}{\partial \tau_i} = \sum_{j=1}^{n} w_{ij} \left[\frac{(y_j - \mu_i)^2}{\tau_i^2} - \frac{1}{2\tau_i} \right] = 0$$

Czyli

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n} w_{ij} (y_j - \hat{\mu}_i)^2}{\sum_{j=1}^{n} w_{ij}}$$

Podsumowanie EM dla mieszanek gaussowskich

Krok E:

$$w_{ij}^{[t+1]} = \frac{\alpha_i^{[t]} \phi_{\mu_i^{[t]}, \sigma_i^{[t]}}(y_j)}{\sum_{i=1}^{k} \alpha_i^{[t]} \phi_{\mu_i^{[t]}, \sigma_i^{[t]}}(y_j)}$$

Krok M:

$$\alpha_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]}}{n}$$

$$\mu_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]} y_j}{\sum_{j=1}^{n} w_{ij}^{[t+1]}}$$

$$(\sigma_i^2)^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]} (y_j - \mu_i^{[t+1]})^2}{\sum_{j=1}^{n} w_{ij}^{[t+1]}}$$

EM dla wielowymiarowych gaussowskich

Krok E:

$$w_{ij}^{[t+1]} = \frac{\alpha_i^{[t]} \phi_{\mu_i^{[t]}, \Sigma_i^{[t]}}(y_j)}{\sum_{i=1}^{k} \alpha_i^{[t]} \phi_{\mu_i^{[t]}, \Sigma_i^{[t]}}(y_j)}$$

Krok M:

$$\alpha_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]}}{n}$$

$$\mu_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]} y_j}{\sum_{j=1}^{n} w_{ij}^{[t+1]}}$$

$$\Sigma_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]} (y_j - \mu_i^{[t+1]})(y_j - \mu_i^{[t+1]})^T}{\sum_{j=1}^{n} w_{ij}^{[t+1]}}$$

Uwagi: Estymowanie k macierzy kowariancji jest trudne (szczególnie dla większego wymiaru). W takiej sytuacji najczęściej zakłada się, że macierz kowariancji jest wspólna dla wszystkich rozkładów lub zakłada się szczególną postać macierzy kowariancji: $\Sigma_i = \sigma_i^2 I$, Σ_i diagonalna, itp.

EM jako proximal algorytm

Algorytm EM można zapisać w postaci

$$\theta_{\mathrm{k}+1} = \arg\max_{\theta} \mathrm{Q}(\theta, \theta_{\mathrm{k}})$$

Ponieważ

$$\mathrm{Q}(\theta, \theta_{\mathrm{k}}) + \mathrm{H}(\theta_{\mathrm{k}}, \theta_{\mathrm{k}}) = \ell(\theta) - \mathrm{KL}(\mathrm{f}_{\theta_{\mathrm{k}}}(\mathrm{x}|\mathrm{y})||\mathrm{f}_{\theta}(\mathrm{x}|\mathrm{y}))$$

Czyli EM równoważnie można zapisać jako

$$\theta_{\mathrm{k}+1} = \arg\min_{\theta}\{-\ell(\theta) + \mathrm{KL}(\mathrm{f}_{\theta_{\mathrm{k}}}(\mathrm{x}|\mathrm{y})||\mathrm{f}_{\theta}(\mathrm{x}|\mathrm{y}))\}$$

Podsumowanie I

1. Spadek po gradiencie
   1. Wymaga obliczania tylko gradinetu, koszt kroku $\mathcal{O}(\mathrm{dn})$
   2. Dla funkcji wypukłych i L gładkich tempo zbieżności $\mathcal{O}(\mathrm{k}^{-1})$
   3. Przyspieszenie Nesterova poprawia tempo zbieżności do $\mathcal{O}(\mathrm{k}^{-2})$
2. Proximal gradient
   1. Pozwala obliczać funkcję niegładkie o ile obliczenie kroku jest możliwe.
   2. Takie same tempo zbieżności co spadek po gradiencie. Istnieje przypsieszenie Nesterova.
3. ADMM
   1. Również pozwala na niegładkie funkcje.
   2. Często szybszy od proximal gradient w szczególności gdy poszczególne kroki można wyznaczyć analitycznie.
4. EM

Podsumowanie II

1. Pozwala efektywnie obliczać estymatory dla modeli z ukrytymi zmiennymi (brakującymi danymi).
2. Krok E polega na imputacji ukrytych zmiennych.
3. EM jest monotoniczny. Dla funckji wypukłych zbiega nie wolniej niż spadek po gradiencie.