Convex Optimization Theory and Applications

Topic 15 - Quasi-Newton Algorithms

Li Li

Department of Automation
Tsinghua University

Fall, 2009-2021.

15.0. Outline

15.1. Motivation of Quasi-Newton Algorithms

15.2. SR1, BFGS, DFP, Broyden Class

15.1. Motivation of Quasi-Newton Algorithms

smooth convex optimization

$$\min_{x} \ f(x)$$

where f is convex, twice differentable, and $\mathrm{dom}(f) = \mathbb{R}^n$. Recall gradient descent update:

$$x^+ = x - t\nabla f(x)$$

and Newton's method update:

$$x^+ = x - t(\nabla^2 f(x))^{-1}\nabla f(x)$$

- Newton's method has (local) quadratic convergence, versus linear convergence of gradient descent
- But Newton iterations are much more expensive ...

15.1. Motivation of Quasi-Newton Algorithms

Two main steps in Newton iteration:

- Compute Hessian $\nabla^2 f(x)$
- Solve the system $\nabla^2 f(x)s = -\nabla f(x)$

Each of these two steps could be expensive

Quasi-Newton methods repeat updates of the form

$$x^+ = x + ts$$

where direction s is defined by linear system

$$Bs = -\nabla f(x)$$

for some approximation B of $\nabla^2 f(x)$. We want B to be easy to compute, and $Bs = g$ to be easy to solve

15.1. Motivation of Quasi-Newton Algorithms

Quasi-Newton template

Let $x^{(0)} \in \mathbb{R}^n$, $B^{(0)} \succ 0$. For $k = 1, 2, 3, \ldots$, repeat:

1. Solve $B^{(k-1)} s^{(k-1)} = -\nabla f(x^{(k-1)})$
2. Update $x^{(k)} = x^{(k-1)} + t_k s^{(k-1)}$
3. Compute $B^{(k)}$ from $B^{(k-1)}$

Different quasi-Newton methods implement Step 3 differently. As we will see, commonly we can compute $(B^{(k)})^{-1}$ from $(B^{(k-1)})^{-1}$

Basic idea: as $B^{(k-1)}$ already contains info about the Hessian, use suitable matrix update to form $B^{(k)}$

Reasonable requirement for $B^{(k)}$ (motivated by secant method):

$$\nabla f(x^{(k)}) = \nabla f(x^{(k-1)}) + B^{(k)} s^{(k-1)}$$

15.1. Motivation of Quasi-Newton Algorithms

We can equivalently write latter condition as

$$\nabla f(x^+) = \nabla f(x) + B^+ s$$

Letting $y = \nabla f(x^+) - \nabla f(x)$, this becomes

$$B^+ s = y$$

This is called the secant equation

In addition to the secant equation, we want:

- B^+ to be symmetric
- B^+ to be "close" to B
- $B \succ 0 \Rightarrow B^+ \succ 0$

15.2. SR1, BFGS, DFP, Broyden Class

Let's try an update of the form:

$$B^+ = B + auu^T$$

The secant equation $B^+s = y$ yields

$$(au^Ts)u = y - Bs$$

This only holds if u is a multiple of $y - Bs$. Putting $u = y - Bs$, we solve the above, $a = 1/(y - Bs)^Ts$, which leads to

$$B^+ = B + \frac{(y - Bs)(y - Bs)^T}{(y - Bs)^Ts}$$

called the symmetric rank-one (SR1) update

15.2. SR1, BFGS, DFP, Broyden Class

How can we solve $B^+ s^+ = -\nabla f(x^+)$, in order to take next step? In addition to propogating B to B^+, let's propogate inverses, i.e., $C = B^{-1}$ to $C^+ = (B^+)^{-1}$

Sherman-Morrison formula:

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}$$

Thus for the SR1 update the inverse is also easily updated:

$$C^+ = C + \frac{(s - Cy)(s - Cy)^T}{(s - Cy)^T y}$$

In general, SR1 is simple and cheap, but has key shortcoming: it does not preserve positive definiteness

15.2. SR1, BFGS, DFP, Broyden Class

Let's now try a rank-two update:

$$B^+ = B + auu^T + bvv^T.$$

The secant equation $y = B^+ s$ yields

$$y - Bs = (au^T s)u + (bv^T s)v$$

Putting $u = y,\ v = Bs$, and solving for a, b we get

$$B^+ = B - \frac{Bss^T B}{s^T Bs} + \frac{yy^T}{y^T s}$$

called the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update

15.2. SR1, BFGS, DFP, Broyden Class

Woodbury formula (generalization of Sherman-Morrison):

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Applied to our case, we get a rank-two update on inverse C:

$$\begin{aligned} C^{+} &= C + \frac{(s - Cy)s^T}{y^T s} + \frac{s(s - Cy)^T}{y^T s} - \frac{(s - Cy)^T y}{(y^T s)^2} ss^T \\ &= \left(I - \frac{sy^T}{y^T s}\right) C \left(I - \frac{ys^T}{y^T s}\right) + \frac{ss^T}{y^T s} \end{aligned}$$

The BFGS update is thus still quite cheap: $O(n^2)$ operations

15.2. SR1, BFGS, DFP, Broyden Class

The Sherman-Morrison-Woodbury formula is

$$(\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} = \mathbf{P}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}$$

Expand the left hand side:

$$\begin{aligned}
& (\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} \\
= \; & (\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}[\mathbf{R}(\mathbf{H}^{\mathrm{T}})^{-1}]^{-1} \\
= \; & [(\mathbf{R}(\mathbf{H}^{\mathrm{T}})^{-1})(\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})]^{-1} \\
= \; & [(\mathbf{R}(\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{P}^{-1} + \mathbf{H}]^{-1} \\
= \; & [(\mathbf{R}(\mathbf{H}^{\mathrm{T}})^{-1} + \mathbf{H}\mathbf{P})\mathbf{P}^{-1}]^{-1} \\
= \; & [(\mathbf{R} + \mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}})(\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{P}^{-1}]^{-1} \\
= \; & \mathbf{P}\mathbf{H}^{\mathrm{T}}(\mathbf{R} + \mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}})^{-1}
\end{aligned}$$

15.2. SR1, BFGS, DFP, Broyden Class

Importantly, BFGS update preserves positive definiteness. Recall this means $B \succ 0 \Rightarrow B^+ \succ 0$. Equivalently, $C \succ 0 \Rightarrow C^+ \succ 0$

To see this, compute

$$x^T C^+ x = \left(x - \frac{s^T x}{y^T s} y\right)^T C \left(x - \frac{s^T x}{y^T s} y\right) + \frac{(s^T x)^2}{y^T s}$$

Now observe:

- Both terms are nonnegative
- Second term is only zero when $s^T x = 0$
- In that case first term is only zero when $x = 0$

15.2. SR1, BFGS, DFP, Broyden Class

Broyden, Fletcher, Goldfarb, Shanno

15.2. SR1, BFGS, DFP, Broyden Class

We could have pursued same idea to update inverse C:

$$C^+ = C + auu^T + bvv^T.$$

Multiplying by y, using the secant equation $s = C^+ y$, and solving for a, b, yields

$$C^+ = C - \frac{Cyy^TC}{y^TCy} + \frac{ss^T}{y^Ts}$$

Woodbury then shows

$$B^+ = \left(I - \frac{ys^T}{y^Ts}\right)B\left(I - \frac{sy^T}{y^Ts}\right) + \frac{yy^T}{y^Ts}$$

This is the Davidon-Fletcher-Powell (DFP) update. Also cheap: $O(n^2)$, preserves positive definiteness. Not as popular as BFGS

15.2. SR1, BFGS, DFP, Broyden Class

Observe that $B^+ \succ 0$ and $B^+ s = y$ imply

$$y^T s = s^T B^+ s > 0.$$

called the curvature condition. Fact: if $y^T s > 0$, then there exists $M \succ 0$ such that $Ms = y$

Interesting alternate derivation for DFP update: find B^+ "closest" to B w.r.t. appropriate conditions, i.e., solve

$$\begin{array}{ll} \min\limits_{B^+} & \|W^{-1}(B^+ - B)W^{-T}\|_F \\ \text{subject to} & B^+ = (B^+)^T \\ & B^+ s = y \end{array}$$

where W is nonsingular and such that $WW^T s = y$. And BFGS solves same problem but with roles of B and C exchanged

15.2. SR1, BFGS, DFP, Broyden Class

SR1, DFP, and BFGS are some of numerous possible quasi-Newton updates. The Broyden class of updates is defined by:

$$B^+ = (1-\phi)B^+_{\text{BFGS}} + \phi B^+_{\text{DFP}}, \quad \phi \in \mathbb{R}$$

By putting $v = y/(y^T s) - Bs/(s^T Bs)$, we can rewrite the above as

$$B^+ = B - \frac{Bss^T B}{s^T Bs} + \frac{yy^T}{y^T s} + \phi(s^T Bs)vv^T$$

Note:

- BFGS corresponds to $\phi = 0$
- DFS corresponds to $\phi = 1$
- SR1 corresponds to $\phi = y^T s/(y^T s - s^T Bs)$

15.2. SR1, BFGS, DFP, Broyden Class

Quasi-Newton iterations for optimization

The field was launched between 1959 and 1970.

William Davidon 1927-
1954 PhD in Physics, U. Chicago
1959: "variable metric" report at Argonne National Lab.
(It was finally published in 1991, first issue of *SIOPT*)
1961-1991: Prof. of Physics and Maths, Haverford Coll

Michael Powell 1936-
1959-1976 Harwell A.E.R.E.
1976- DAMTP, U. of Cambridge
1983 FRS

Charles Broyden 1933-2011
1955-1965: English Electric
1965: "good" and "bad" Broyden methods
1967-1986 U. of Essex
1990-2003 U. of Bologna

Roger Fletcher 1939-
1969-1973 Harwell A.E.R.E.... U. of Leeds
1963: Davidon-Fletcher-Powell paper
1971-2005 U. of Dundee
2003 FRS

15.2. SR1, BFGS, DFP, Broyden Class

Assume that f convex, twice differentiable, having $\mathrm{dom}(f) = \mathbb{R}^n$, and additionally

- ∇f is Lipschitz with parameter L
- f is strongly convex with parameter m
- $\nabla^2 f$ is Lipschitz with parameter M

(same conditions as in the analysis of Newton's method)

Theorem: Both DFP and BFGS, with backtracking line search, converge globally. Furthermore, for all $k \geq k_0$,

$$\|x^{(k)} - x^\star\|_2 \leq c_k \|x^{(k-1)} - x^\star\|_2$$

where $c_k \to 0$ as $k \to \infty$. Here k_0, c_k depend on L, m, M

This is called local superlinear convergence

15.2. SR1, BFGS, DFP, Broyden Class

复杂的来源是, H_k 与 x_k 会互相影响, 每一步所用到的 H_k 既包含 x_{k-1} 附近函数的Hessian信息, 又包含了更早的 $x_{k-2}, x_{k-3}, \ldots$ 附近Hessian的信息, 使得分析起来异常的困难. 上面这篇89年的paper是采用分三步走的方法.

首先, 注意到 $H_k \nabla f(x_k)$ 的方向和 $\nabla f(x_k)$ 是不同的, 这两个方向成一个锐角 θ_k (因为矩阵 H_k 正定). 优化里的一个核心结果是, 要证明 $\{x_k\}$ 是收敛的, 必须证明 θ_k 不能过于接近 $\pi/2$, 或者说 $\cos\theta_k$ 不能太接近于0. 因此, 基于假设

$$\mu I \preceq \nabla^2 f(x) \preceq LI, \quad 0 < \mu \leq L$$

作者们首先证明了, 对于任意 $p \in (0,1)$, 存在一个**不依赖**于 K 的bound $\Delta > 0$, 使得在前 K 个迭代点 $\{x_0, x_1, \cdots, x_{K-1}\}$ 中至少有 $\lceil pK \rceil$ 个满足 $\cos\theta_k \geq \Delta$. 神奇的是, 并不能推广到 $p = 1$ 的情形; 也就是说总有一些"坏"的迭代点是不能被bounded的.

15.2. SR1, BFGS, DFP, Broyden Class

然后, 尽管如此, 作者们还是能通过这一个性质来证明, $\{x_0, x_1, \cdots, x_k, \cdots\}$ 线性地收敛到最优点 x^* .

最后, **因为**这些点 $\{x_0, x_1, \cdots, x_k, \cdots\}$ 线性地收敛到最优点 x^* , 可以注意到当 k 很大时, $\|x_{k+1} - x_k\|$ 已经非常小, 而 x_k, x_{k+1} 都很靠近最优点 x^* , 假定Hessian本身是充分连续的, 那么我们基本上就有

$$\nabla^2 f(x^*)(\nabla f(x_{k+1}) - \nabla f(x_k)) \approx x_{k+1} - x_k$$

也就是说每次我们更新 H_k 时, 收集到的是准确的关于 $\nabla^2 f(x^*)$ 的信息, 因此最终作者们证明了

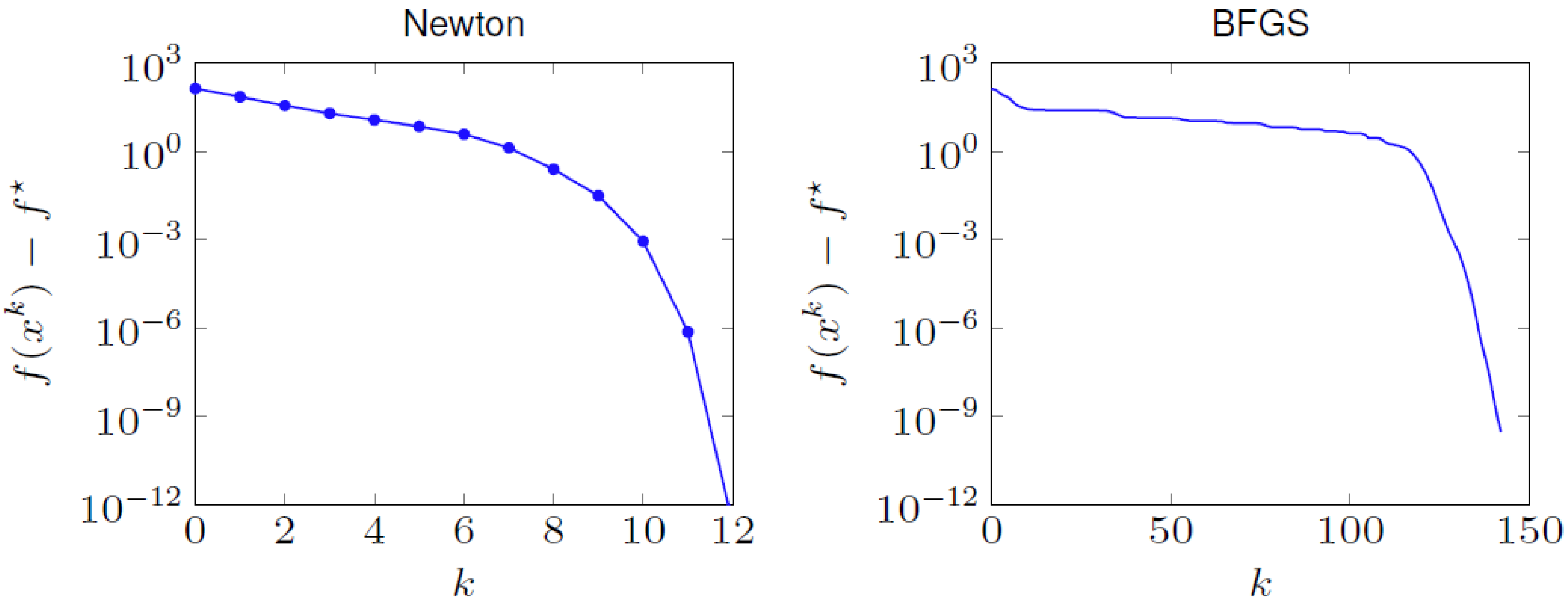
$$\lim_{k\to\infty} \frac{\|(\nabla^2 f(x^*) - H_k^{-1})s_k\|}{\|s_k\|} = 0$$

这一点再加上其他条件还可以推出超线性收敛

15.2. SR1, BFGS, DFP, Broyden Class

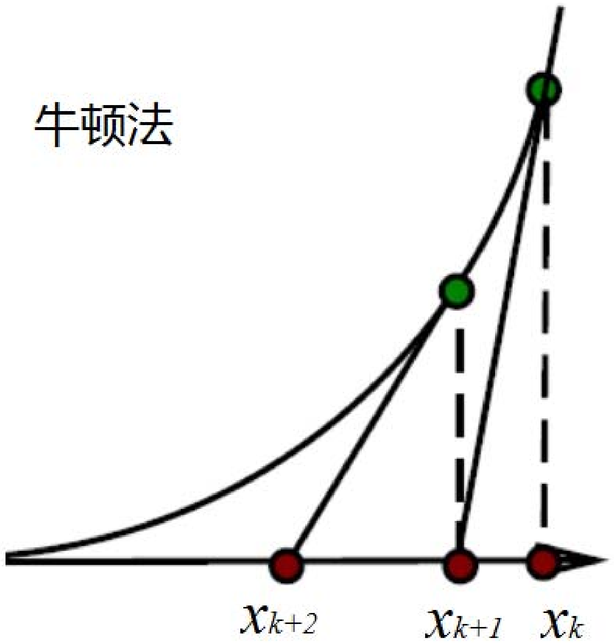
Example from Vandenberghe's lecture notes: Newton versus BFGS on LP barrier problem, for $n = 100$, $m = 500$

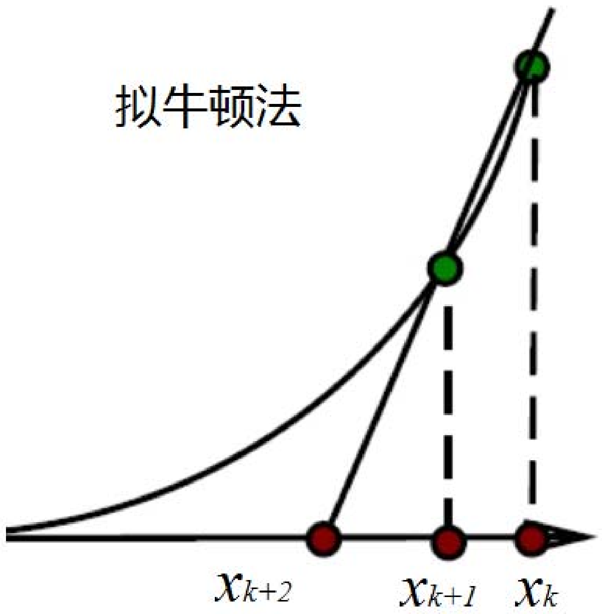
$$\min_x \; c^T x - \sum_{i=1}^m \log(b_i - a_i^T x)$$



Note that Newton update is $O(n^3)$, quasi-Newton update is $O(n^2)$. But quasi-Newton converges in less than 100 times the iterations

15.2. SR1, BFGS, DFP, Broyden Class





这张图是某一个函数的**导函数的图像**。在这种情况下，我们可以看出，如果在导函数上某一点做切线，这条切线的斜率就是二次导函数，并且对应的就是下面这个式子

> $$f''(x_k) = -\frac{f'(x_k)}{x_{k+1} - x_k}$$

化简一下，就是 $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$ ，这个就是牛顿法的一元形式。

考虑 导函数之前两个点所形成的割 线，那么这个时候会得到

> $$B_k(x_k - x_{k-1}) = f'(x_k) - f'(x_{k-1})$$

这里的 B_k 就是**这条割线的斜率**。所以割线法其实就是拟牛顿法的前身，因为如果我们设 $s_k = x_{k+1} - x_k$ ，$y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ ，式子就会变成

> $$B_k s_{k-1} = y_{k-1}$$

这就是拟牛顿法的本质。

15.2. SR1, BFGS, DFP, Broyden Class

For large problems, quasi-Newton updates can become too costly

Basic idea: instead of explicitly computing and storing C, compute an implicit version of C by maintaining all pairs (y, s)

Recall BFGS updates C via

$$C^+ = \left(I - \frac{sy^T}{y^T s}\right) C \left(I - \frac{ys^T}{y^T s}\right) + \frac{ss^T}{y^T s}$$

Observe this leads to

$$C^+ g = p + (\alpha - \beta)s, \quad \text{where}$$
$$\alpha = \frac{s^T g}{y^T s},\ q = g - \alpha y,\ p = Cq,\ \beta = \frac{y^T p}{y^T s}$$

15.2. SR1, BFGS, DFP, Broyden Class

We see that $C^{+}g$ can be computed via too loops of length k (if C^{+} is the approximation to the inverse Hessian after k iterations):

1. Let $q = -\nabla f(x^{(k)})$
2. For $i = k-1, \ldots, 0$:
   (a) Compute $\alpha_i = (s^{(i)})^T q / ((y^{(i)})^T s^{(i)})$
   (b) Update $q = q - \alpha y^i$
3. Let $p = C^{(0)} q$
4. For $i = 0, \ldots, k-1$:
   (a) Compute $\beta = (y^{(i)})^T p / ((y^{(i)})^T s^{(i)})$
   (b) Update $p = p + (\alpha_i - \beta) s^{(i)}$
5. Return p

15.2. SR1, BFGS, DFP, Broyden Class

Limited memory BFGS (LBFGS) simply limits each of these loops to be length m:

1. Let $q = -\nabla f(x^{(k)})$
2. For $i = k-1, \ldots, k-m$:
   (a) Compute $\alpha_i = (s^{(i)})^T q/((y^{(i)})^T s^{(i)})$
   (b) Update $q = q - \alpha y^i$
3. Let $p = \bar{C}^{(k-m)} q$
4. For $i = k-m, \ldots, k-1$:
   (a) Compute $\beta = (y^{(i)})^T p/((y^{(i)})^T s^{(i)})$
   (b) Update $p = p + (\alpha_i - \beta) s^{(i)}$
5. Return p

In Step 3, $\bar{C}^{(k-m)}$ is our guess at $C^{(k-m)}$ (which is not stored). A popular choice is $\bar{C}^{(k-m)} = I$, more sophisticated choices exist

15.3. References

[1] http://www.stat.cmu.edu/~ryantibs/convexopt/lectures/quasi-newton.pdf

[2] http://fourier.eng.hmc.edu/e176/lectures/algebra/node6.html

[3] https://www.caam.rice.edu/~dennis/images/Broyden-Fletcher-Goldfarb-Shanno.jpg

[4] https://www.slideshare.net/MaryamShamsSolary/inventorstalk-62781058

[5] R. H. Byrd, J. Nocedal, "A tool for the analysis of quasi-Newton methods with application to unconstrained minimization," *SIAM Journal on Numerical Analysis*, vol. 26, no. 3, pp. 727-739, 1989.

[6] J. E. Dennis, J. J. Moré, "Application to quasi-Newton methods," *Mathematics of Computation*," vol. 28, no. 126, pp. 549-560, 1974.

[7] https://www.zhihu.com/question/27940474/answer/605933345

[8] https://zhuanlan.zhihu.com/p/144736223