ЛЕКЦИЯ 6
Глава 3. МЕТОДЫ ДЛЯ ЗАДАЧ БЕЗУСЛОВНОЙ ОПТИМИЗАЦИИ И НЕЛИНЕЙНЫХ УРАВНЕНИЙ
(завершение)

Содержание лекции

1. Метод Ньютона. Квазиньютоновские методы
   - Метод Ньютона для уравнений
   - Метод Ньютона для задачи безусловной оптимизации
   - Квазиньютоновские методы

Ньютоновские методы

Фундаментальная идея метода Ньютона чрезвычайно важна: она лежит в основе многих (если не большинства) практических алгоритмов для различных классов вариационных задач и задач оптимизации.
В отличие от методов спуска, в своих базовых формах метод Ньютона — не оптимизационный, а вариационный: он одинаково хорошо локально сходится не только к решениям задачи оптимизации, но к любым ее стационарным точкам.
С другой стороны, ньютоновские методы допускают оптимизационные трактовки, дающие, в частности, почву для глобализации их сходимости.

Ньютоновские методы

Высокая скорость сходимости метода Ньютона достигается за счет того, что для задач оптимизации он является методом второго порядка, и его итерация существенно более трудоемка, чем у градиентных методов.
К счастью, на основе метода Ньютона строятся так называемые квазиньютоновские методы, которые сохраняют высокую скорость сходимости метода Нтютона, а их итерации лишь немногим более трудоемки, чем итерации градиентных методов.

Метод Ньютона для уравнений

Уравнение

$$\Phi(x) = 0, \tag{1}$$

где $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ — гладкое отображение.

Итерационное уравнение *метода Ньютона*

Получается линеаризацией уравнения (1) в текущем приближении $x^k \in \mathbb{R}^n$:

$$\Phi(x^k) + \Phi'(x^k)(x - x^k) = 0. \tag{2}$$

Метод Ньютона для уравнений

Итерационная схема метода Ньютона

$$x^{k+1} = x^k - (\Phi'(x^k))^{-1}\Phi(x^k), \quad k = 0, 1, \ldots$$

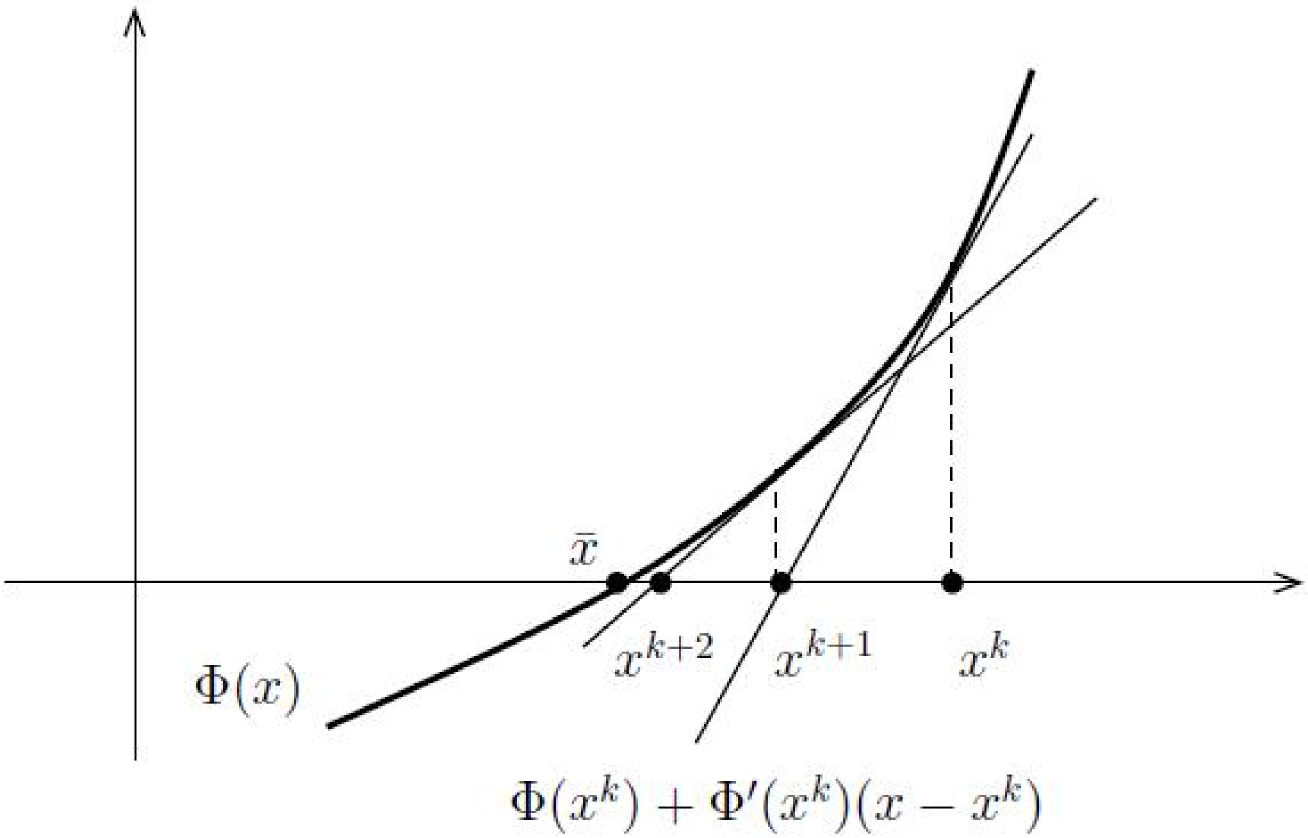
Подразумевает невырожденность $\Phi'(x^k)$ для всех k, хотя практическое обращение этих матриц не требуется.

Алгоритм 1

Выбираем $x^0 \in \mathbb{R}^n$ и полагаем $k = 0$.

1. Вычисляем $x^{k+1} \in \mathbb{R}^n$ как решение уравнения (2).
2. Увеличиваем номер шага k на 1 и переходим к п. 1.

Метод Ньютона для уравнений



Локальная сверхлинейная сходимость

Теорема 1

Пусть отображение $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ дифференцируемо в некоторой окрестности точки $\bar{x} \in \mathbb{R}^n$, причем его производная непрерывна в этой точке. Пусть $\bar{x}$ является решением уравнения (1), причем $\det \Phi'(\bar{x}) \neq 0$.
Тогда любое начальное приближение $x^0 \in \mathbb{R}^n$, достаточно близкое к $\bar{x}$, корректно определяет траекторию алгоритма 1, которая сходится к $\bar{x}$.
Скорость сходимости сверхлинейная, а если производная Φ липшицева относительно $\bar{x}$, т.е. для $x \in \mathbb{R}^n$

$$\Phi'(x) - \Phi'(\bar{x}) = O(\|x - \bar{x}\|),$$

то квадратичная.

Метод Ньютона для уравнений

Доказательство

Из линейной алгебры и/или анализа: найдутся числа $\delta > 0$ и $M > 0$ такие, что для всякого $x \in B(\bar{x}, \delta)$ матрица $\Phi'(x)$ невырождена и $\|(\Phi'(x))^{-1}\| \leqslant M$.
Тогда для $x^k \in B(\bar{x}, \delta)$ следующее приближение x^{k+1} корректно определено и, в силу теоремы о среднем,

$$\|x^{k+1} - \bar{x}\| = \|x^k - \bar{x} - (\Phi'(x^k))^{-1}\Phi(x^k)\| \leqslant$$

$$\leqslant \|(\Phi'(x^k))^{-1}\|\|\Phi(x^k) - \Phi(\bar{x}) - \Phi'(x^k)(x^k - \bar{x})\| \leqslant$$

$$\leqslant M \sup_{t\in[0,1]} \|\Phi'(tx^k + (1-t)\bar{x}) - \Phi'(x^k)\|\|x^k - \bar{x}\| = o(\|x^k - \bar{x}\|).$$

В частности, для любого $q \in (0, 1)$, если δ достаточно мало, то $\|x^{k+1} - \bar{x}\| \leqslant q\|x^k - \bar{x}\|$, и поэтому $x^{k+1} \in B(\bar{x}, \delta)$.

Метод Ньютона для уравнений

Доказательство (продолжение)

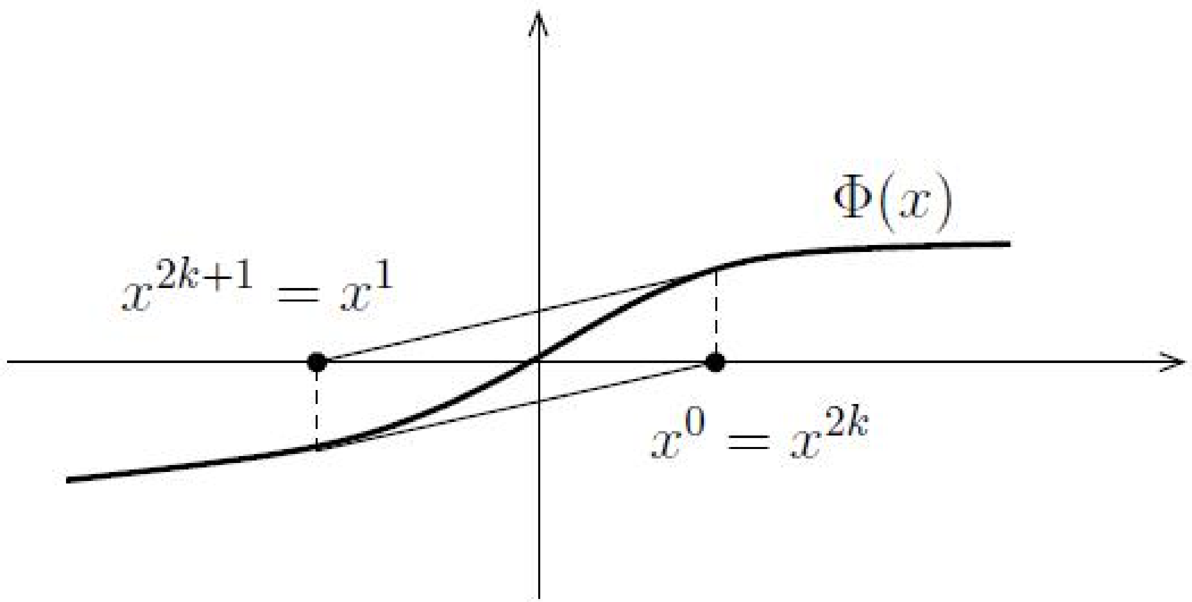
Из сказанного следует требуемое.
В частности, если производная Φ липшицева относительно $\bar{x}$, то из оценки выше

$$x^{k+1} - \bar{x} = O(\|x^k - \bar{x}\|^2),$$

что дает квадратичную скорость сходимости.

Метод Ньютона обладает лишь локальной сходимостью!

Метод Ньютона для уравнений



Метод Ньютона для задачи безусловной оптимизации

Задача

$$f(x) \to \min, \quad x \in \mathbb{R}^n, \tag{3}$$

где $f : \mathbb{R}^n \to \mathbb{R}$ — заданная дважды дифференцируемая функция.

Стационарные точки характеризуются уравнением (1) с градиентным отображением

$$\Phi : \mathbb{R}^n \to \mathbb{R}^n, \quad \Phi(x) = f'(x).$$

Итерационная схема метода Ньютона

$$x^{k+1} = x^k - (f''(x^k))^{-1} f'(x^k), \quad k = 0, 1, \ldots \tag{4}$$

В частности, это метод второго порядка.

Метод Ньютона для задачи безусловной оптимизации

Итерационное уравнение метода Ньютона в текущем приближении $x^k \in \mathbb{R}^n$

$$f'(x^k) + f''(x^k)(x - x^k) = 0. \tag{5}$$

Оптимизационная трактовка *метода Ньютона* для задачи безусловной оптимизации

Уравнение (5) характеризует стационарные точки задачи

$$f(x^k) + \langle f'(x^k), x - x^k \rangle +$$

$$+ \frac{1}{2}\langle f''(x^k)(x - x^k), x - x^k \rangle \to \min, \quad x \in \mathbb{R}^n, \tag{6}$$

целевая функция которой есть квадратичная аппроксимация f вблизи x^k.

Метод Ньютона для задачи безусловной оптимизации

Алгоритм 2

Выбираем $x^0 \in \mathbb{R}^n$ и полагаем $k = 0$.

1. Вычисляем $x^{k+1} \in \mathbb{R}^n$ как стационарную точку задачи (6).
2. Увеличиваем номер шага k на 1 и переходим к п. 1.

Локальная сверхлинейная сходимость

Следствие 1

Пусть функция $f : \mathbb{R}^n \to \mathbb{R}$ дважды дифференцируема в некоторой окрестности точки $\bar{x} \in \mathbb{R}^n$, причем ее вторая производная непрерывна в этой точке. Пусть $\bar{x}$ — стационарная точка задачи (3), причем в этой точке выполнено достаточное условие второго порядка оптимальности (т.е. матрица $f''(\bar{x})$ положительно определена).

Тогда любое начальное приближение $x^0 \in \mathbb{R}^n$, достаточно близкое к $\bar{x}$, корректно определяет траекторию алгоритма 2, которая сходится к $\bar{x}$.

Скорость сходимости сверхлинейная, а если вторая производная f липшицева относительно $\bar{x}$, то квадратичная.

Метод Ньютона для задачи безусловной оптимизации

Оптимизационная специфика: матрица Гессе дважды дифференцируемой функции симметрична.
В предположениях следствия 1 рациональный метод решения итерационного линейного уравнения (5) — алгоритм Холецкого (квадратного корня): дает $LL^\top$-разложение положительно определенной симметричной матрицы за порядка $n^3/6$ операций.
Часто: при нарушении положительной определенности $f''(x^k)$ (вдали от $\bar{x}$) ее заменяют на $f''(x^k) + \gamma I$ при достаточно большом $\gamma > 0$, либо на другую положительно определенную матрицу. Последнее можно реализовать как модификацию алгоритма Холецкого. Важно для глобализации сходимости.

Без модификаций метод Ньютона «не отличает» минимумы от других стационарных точек: в следствии 1 достаточное условие оптимальности можно заменить на $\det f''(\bar{x}) \neq 0$.

Достоинства и недостатки

Достоинство метода Ньютона: сверхлинейная скорость сходимости (в отличие от градиентных методов).

Недостатки метода Ньютона:

- только локальная сходимость (в отличие от градиентных методов);
- существенно более трудоемкая итерация (по сравнению с градиентными методами): нужно вычислять матрицу Гессе целевой функции и решать линейную систему.

Достоинства и недостатки

Для квадратичной функции

$$f(x) = \langle Ax, x\rangle + \langle b, x\rangle, \quad x \in \mathbb{R}^n, \tag{7}$$

где $A \in \mathbb{R}(n, n)$ — невырожденная симметричная матрица, $b \in \mathbb{R}^n$, метод Ньютона находит единственную стационарную точку из любого начального приближения (за один шаг).

В неквадратичном случае при неудачном начальном приближении предельные точки траектории метода могут не быть стационарными.

Как избавиться от недостатков?

Стратегии глобализации сходимости ньютоновских методов используют:

- оптимизационную трактовку;
- параметр длины шага:

$$x^{k+1} = x^k - \alpha_k (f''(x^k))^{-1} f'(x^k), \quad k = 0, 1, \ldots,$$

где $\alpha_k > 0$ может быть меньше 1 (возможно в сочетании с заменой $f''(x^k)$ ее положительно определенной модификацией, что делает эту схему методом спуска).

Далее: средства удешевления итерации с сохранением высокой скорости сходимости.

Квазиньютоновские методы

Класс методов спуска

$$x^{k+1} = x^k + \alpha_k d^k, \quad d^k = -Q_k f'(x^k), \quad k = 0, 1, \dots, \tag{8}$$

где $Q_k \in \mathbb{R}(n, n)$ — симметричная положительно определенная матрица, $\alpha_k > 0$ — параметр длины шага.

Если $f'(x^k) \neq 0$, то

$$\langle f'(x^k), d^k \rangle = -\langle Q_k f'(x^k), f'(x^k) \rangle < 0, \tag{9}$$

и в силу леммы 3.1.1 $d^k \in \mathcal{D}_f(x^k)$.

Частные случаи (8):

- градиентные методы ($Q_k = I$);
- метод Ньютона ($Q_k = (f''(x^k))^{-1}$ и $\alpha_k = 1$).

Квазиньютоновские методы

Алгоритм 3

Выбираем $x^0 \in \mathbb{R}^n$ и полагаем $k = 0$. Выбираем одно из трех основных правил одномерного поиска и необходимые для реализации этого правила параметры.

1. Выбираем симметричную положительно определенную матрицу $Q_k \in \mathbb{R}(n, n)$. Полагаем $d^k = -Q_k f'(x^k)$ и вычисляем α_k в соответствии с выбранным правилом одномерного поиска по направлению d^k.
2. Полагаем $x^{k+1} = x^k + \alpha_k d^k$.
3. Увеличиваем номер шага k на 1 и переходим к п. 1.

Квазиньютоновские методы

Квазиньютоновские методы

Это методы спуска класса (8), в которых Q_k аппроксимируют $(f''(\bar{x}))^{-1}$ в искомом решении $\bar{x}$ в смысле условия (10) в следующей теореме.

Теорема Дэнниса–Морэ

Теорема 2 (Дэннис, Морэ)

Пусть выполнены предположения следствия 1. Пусть, кроме того, траектория $\{x^k\}$ алгоритма 3, использующего правило Армихо при $\hat{\alpha} = 1$ и $\varepsilon \in (0, 1/2)$, сходится к $\bar{x}$.
Тогда следующие два утверждения эквивалентны:

- скорость сходимости последовательности $\{x^k\}$ к $\bar{x}$ является сверхлинейной, причем $\alpha_k = 1$ для любого достаточно большого k;
- имеет место предельное соотношение

$$(Q_k - (f''(\bar{x}))^{-1})f'(x^k) = o(\|f'(x^k)\|). \qquad (10)$$

Теорема Дэнниса–Морэ

Соотношение Дэнниса–Морэ (10) равносильно

$$(Q_k^{-1} - f''(\bar{x}))Q_k f'(x^k) = o(\|Q_k f'(x^k)\|). \tag{11}$$

Вывод: для построения эффективных численных методов оптимизации *необходимо* использовать информацию «второго порядка». Это возможно без фактического вычисления вторых производных.

Квазиньютоновское уравнение

Идея квазиньютоновских методов: заменить вычисление $f''(x^k)$ и решение итерационной линейной системы прямой аппроксимацией $(f''(\bar{x}))^{-1}$ в смысле соотношения Дэнниса–Морэ.

Положим

$$r^k = x^{k+1} - x^k, \quad s^k = f'(x^{k+1}) - f'(x^k) \tag{12}$$

(известны к тому моменту, когда нужно вычислять Q_{k+1}). Формализация идеи — *квазиньютоновское уравнение*:

$$Q_{k+1} s^k = r^k. \tag{13}$$

Квазиньютоновское уравнение

Действительно, при $\alpha_k = 1$ из (8) и (12) имеем $r^k = -Q_k f'(x^k)$. Стремление к выполнению условия Дэнниса–Морэ (11) можно интерпретировать так: нужно выбирать Q_{k+1} таким образом, чтобы вектор $Q_{k+1}^{-1} r^k = -Q_{k+1}^{-1} Q_k f'(x^k)$ был как можно ближе к $f''(\bar{x}) r^k = -f''(\bar{x}) Q_k f'(x^k)$.
По формуле Ньютона–Лейбница

$$s^k = f'(x^{k+1}) - f'(x^k) = \int_0^1 f''(tx^{k+1} + (1-t)x^k) r^k \, dt \approx f''(\bar{x}) r^k$$

при x^k и x^{k+1}, близких к $\bar{x}$. Поэтому естественно требовать $Q_{k+1}^{-1} r^k = s^k$, а это равносильно квазиньютоновскому уравнению (13).

Квазиньютоновские методы

Естественное дополнительное к квазиньютоновскому уравнению (13) требование: «минимальность» поправки $Q_{k+1} - Q_k$ (устойчивость: от итерации к итерации Q_k должны меняться как можно меньше).
Конкретное понимание «минимальности» приводит к различным квазиньютоновским методам.

Квазиньютоновские методы

Метод Давидона–Флетчера–Пауэлла (ДФП)

Q_0 — произвольная положительно определенная симметричная матрица (например, $Q_0 = I$),

$$Q_{k+1} = Q_k + \frac{r^k (r^k)^\top}{\langle r^k,\, s^k \rangle} - \frac{(Q_k s^k)(Q_k s^k)^\top}{\langle Q_k s^k,\, s^k \rangle}. \tag{14}$$

Исторически первый квазиньютоновский метод.

Квазиньютоновские методы

Матрицы ДФП симметричные и удовлетворяют квазиньютоновскому уравнению (13). Действительно,

$$Q_{k+1}s^k = Q_k s^k + r^k \frac{\langle r^k,\, s^k\rangle}{\langle r^k,\, s^k\rangle} - Q_k s^k \frac{\langle Q_k s^k,\, s^k\rangle}{\langle Q_k s^k,\, s^k\rangle} =$$

$$= Q_k s^k + r^k - Q_k s^k = r^k.$$

Ранг поправки $Q_{k+1} - Q_k$ не превосходит 2: $\ker(Q_{k+1} - Q_k)$ содержит все векторы, ортогональные r^k и $Q_k s^k$, и поправка «мала» по крайней мере в этом смысле.

Квазиньютоновские методы

Положительная определенность матриц ДФП зависит от способа выбора параметра длины шага.

Предложение 1

Пусть $Q_k \in \mathbb{R}(n, n)$ — симметричная положительно определенная матрица. Пусть функция $f : \mathbb{R}^n \to \mathbb{R}$ дифференцируема в точках x^k, $x^{k+1} \in \mathbb{R}^n$, которые связаны формулой (8) с некоторым параметром длины шага $\alpha_k > 0$. Тогда формулы (12) и (14) корректно определяют положительно определенную матрицу Q_{k+1} в том и только том случае, когда выполняется неравенство

$$\langle f'(x^{k+1}), d^k \rangle > \langle f'(x^k), d^k \rangle. \tag{15}$$

Квазиньютоновские методы

Условие (15) выполняется, если параметр длины шага выбирается по правилу:

- одномерной минимизации ($\langle f'(x^{k+1}), d^k \rangle = 0$ согласно (3.1.5), и из (9) следует (15));
- *Вулфа*.

Правила Армихо и Голдстейна этим свойством не обладают.

Квазиньютоновские методы

Метод Бройдена–Флетчера–Голдфарба–Шэнно (БФГШ)

Q_0 — произвольная положительно определенная симметричная матрица (например, $Q_0 = I$),

$$Q_{k+1} = Q_k + \frac{(r^k - Q_k s^k)(r^k)^\top + r^k(r^k - Q_k s^k)^\top}{\langle r^k, s^k \rangle} - \\ - \frac{\langle r^k - Q_k s^k, s^k \rangle r^k (r^k)^\top}{\langle r^k, s^k \rangle^2}. \qquad (16)$$

Теоретические свойства аналогичны ДФП.

Признается наиболее эффективным из известных квазиньютоновских методов.

Квазиньютоновские методы

ДФП и БФГШ «взаимодвойственны»: для произвольной симметричной положительно определенной матрицы $Q_k \in \mathbb{R}(n, n)$ и для $H_k = Q_k^{-1}$, если Q_{k+1} сгенерирована по формуле (16) БФГШ, а H_{k+1} — согласно формуле (14) ДФП, причем H_{k+1} невырождена, то и Q_{k+1} невырождена, причем $H_{k+1} = Q_{k+1}^{-1}$.
Отсюда следует, что аналог предложения 1 справедлив и для БФГШ.

Для квадратичной функции вида (7) с положительно определенной матрицей A методы ДФП и БФГШ с правилом одномерной минимизации находят единственную стационарную точку (по необходимости являющуюся глобальным решением задачи (3)) из любого начального приближения за $\bar{k} \leqslant n$ шагов, причем $Q_{\bar{k}} = (f''(x^{\bar{k}}))^{-1} = A^{-1}/2$.

Квазиньютоновские методы

В неквадратичном случае анализ сходимости и скорости сходимости методов ДФП и БФГШ крайне нетривиален. Известные результаты о глобальной сходимости для БФГШ с правилом Вулфа и для ДФП с правилом одномерной минимизации предполагают выпуклость f.

Доказательство сверхлинейной скорости сходимости для конкретных квазиньютоновских методов сводится к проверке условия Дэнниса–Морэ и применению теоремы Дэнниса–Морэ.

Квазиньютоновские методы

Именно появление квазиньютоновских методов сделало численную оптимизацию практической наукой, позволив за приемлемое время решать большие прикладные задачи.
И по сей день квазиньютоновские методы остаются основным практическим средством численного решения задач безусловной оптимизации, поскольку сочетают высокую скорость сходимости с невысокой трудоемкостью итерации.

С теоретической точки зрения любой квазиньютоновский метод есть не более чем умная реализация фундаментальной идеи метода Ньютона, однако, это тот случай, когда реализация имеет не меньшее значение, чем реализуемая идея.

Принципы квазиньютоновских методов распространяются и на методы для задач условной оптимизации и вариационных задач.