0.1 Второй подход, многомерная дихотомия

Рассмотрим задачу минимизации функции на гиперкубе:

$$\min_{\mathbf{x}\in Q} f(x, y),$$

где f - это выпуклая функция, $Q = [0, a]^n \subset \mathbb{R}^n, a > 0$ - гиперкуб. Здесь и далее будем считать, что стороны квадрата соориентированы параллельно осям текущей системы координат. Очевидно, что это не предположение не существенно, поскольку для любого гиперкуб существует тривиальное афинное преобразование, приводящее его к такому виду.

Тогда рассмотрим метод многомерные дихотомии. Его основные этапы следующие:

1. Для всех осей Ox_k рассмотреть гиперкуб $Q_k = \left\{\mathbf{x} \in Q \middle| x_k = \frac{\max_{\mathbf{y}\in Q} y_k + \min_{\mathbf{y}\in Q} y_k}{2}\right\}$.

2. Решить задачу минимизации f на Q_k с точностью Δ (по функции или по аргументу)при помощи этого метода для Q_k

3. Взять δ-субградиент в этой точке и оставить ту часть, в котору смотрит антиградиент.

4. Повторять шаги 1-3 до достижения необходимой точности

Заметим, что когда $\dim_Q = 1$, то полученное на первом шаге Q_k содержит только одну точку.

Сейчас обсудим корректность алгоритма. Здесь и далее мы будем использовать следующую нотацию. Пусть на текущем уровне рекурсии рассматривается множество Q и его некоторое сечение Q_k. Пусть $\nabla_\delta f(\mathbf{x})$ - δ-субградиент функции в точке $\mathbf{x}$. Тогда определим $\nabla_{\delta,\|Q_k} f(\mathbf{x})$, $\nabla_{\delta,\perp Q_k} f(\mathbf{x})$ - проекции δ субградиента на Q_k и его ортогональное дополнение в пространстве $\mathbb{R}^{\dim Q}$ соответсвенно.

Здесь и далее нам понадобятся стандартные понятия субдифференциала и субградиента, которые можно найти, например, в [**?**].

Нам понадобится следующая лемма.

Лемма 0.1. *Пусть f есть непрерывно дифференцируемая выпуклая функция. Пусть на гиперкубе Q решается задача минимизации на Q_k. Если $\boldsymbol{x}_*$ решение этой задачи:*

$$\exists g \in \partial_Q f(\boldsymbol{x}_*) : g_{\|} = 0$$

Доказательство. Если $\mathbf{x}_*$ есть внутренняя точка, то градиент по нефиксированным переменнам равен нулю в силу того, что $\mathbf{x}_*$ - минимум. Тогда с учетом того, что $\nabla f(\mathbf{x}_*) \in \partial f(\mathbf{x}_*)$, получаем утвержение из теоремы.

Допустим, что $\mathbf{x}_*$ граничная точка. Тогда множество условного субдифференциала на гиперкубе Q определяется следующим образом:

$$\partial_Q f(\mathbf{x}) = \partial f(\mathbf{x}) + N\left(\mathbf{x}|Q\right),$$

где $N(\mathbf{x}|Q) = \{\mathbf{a}|\langle \mathbf{a}, \mathbf{y} - \mathbf{x}\rangle \leq 0, \forall \mathbf{y} \in Q\}$.

В случае дифференцируемой функции имеем:

$$\partial f(\mathbf{x}_*) = \{\nabla f(\mathbf{x}_*)\},$$

Из того, что $\mathbf{x}_*$ есть внутренняя точка, следует, что существует непустой набор координат $\{x_j\}_j$, такой что $x_j = \max_{\mathbf{y}\in Q_k} y_j$ или $x_j = \min_{\mathbf{y}\in Q_k} y_j$. Введем обозначение:

$$J_+ = \{j \in \mathbb{N}|x_j = \max_{\mathbf{y}\in Q_k} y_j\}$$

$$J_- = \{j \in \mathbb{N}|x_j = \min_{\mathbf{y}\in Q_k} y_j\}$$

В таком случае заметим, что любой вектор a такой, что $a_j \geq 0, \forall j \in J_+$, $a_j \leq 0, \forall j \in J_-$ и $a_j = 0$ в противном случае, принадлежит нормальному конусу.

Также заметим, что $(\nabla f(\mathbf{x}_*))_j \leq 0, \forall j \in J_+$, $(\nabla f(\mathbf{x}_*))_j 0, \forall j \in J_-$ и $(\nabla f(\mathbf{x}_*))_j = 0$. Действительно, если $(\nabla f(\mathbf{x}_*))_j > 0$ для некоторого $j \in J_+$, то существует вектор $\mathbf{x} = \mathbf{x}_* + \alpha \mathbf{e}_k \in Q$ для некоторого $\alpha < 0$ и вектора $e_k^j = \delta_{kj}$, причем значение в этой функции будет $f(\mathbf{x}) = f(\mathbf{x}_*) + \alpha(\nabla f(\mathbf{x}_*))_j + o(\alpha) < f(\mathbf{x}_*)$ для достаточно малого α, что противоречит тому, что $\mathbf{x}_*$ - решение.

Тогда выбрав $\mathbf{a}$, такой что $\mathbf{a}_\parallel = -\left(\nabla f(\mathbf{x}_*)\right)_\parallel$, получаем субградиент из условия:

$$\mathbf{g} = \nabla f(\mathbf{x}_*) + \mathbf{a} : \mathbf{g}_\parallel = 0$$

□

Данный метод работает не для всех выпуклых функций, даже если решать задачу одномерной оптимизации точно. Пример негладкой выпуклой функции был расмотрен в [2].

Обобщая результат из работы [2], получаем следующую оценку. Если задан гиперкуб Q и требуется минимизировать на нем функцию с точностью

ϵ, то для этого в методе достаточно решить вспомогательную задачу с точностью по аргументу

$$\Delta \leq \frac{\epsilon}{8L_f\sqrt{n}a}, \tag{1}$$

где $n = \dim Q$, R - есть размер начального гиперкуба. Если f есть μ_f-сильно выпуклая функция, то используя выше написанное условие останова, получаем, что достаточно решить вспомогательную задачу с точностью $\tilde{\epsilon}$ по функции:

$$\tilde{\epsilon} \leq \frac{\mu_f \epsilon^2}{128L_f^2 na^2} \tag{2}$$

Теперь допустим, что у нас не доступен вектор $\nabla f(\mathbf{x})$, но доступно его приближение $\nu(\mathbf{x})$, такое что $\|\nabla f(\mathbf{x}) - \nu(\mathbf{x})\| \leq \tilde{\delta}$. Тогда наш метод сойдется к ϵ-решению при использовании $\nu(\mathbf{x})$ вместо точного градиента при выполнении следующего условия

$$L_f\Delta + 2\tilde{\delta} \leq \frac{\epsilon}{4L_f\sqrt{n}a}. \tag{3}$$

Данный результат доказывается аналогично замечанию 1 из работы [2]. В частности:

$$\Delta \leq \frac{\epsilon}{16L_f\sqrt{n}a}, \tag{4}$$

$$\tilde{\delta} \leq \frac{\epsilon}{32\sqrt{n}a}, \tag{5}$$

Тогда для точности по функции решения вспомогательной задачи достаточно

$$\tilde{\epsilon} \leq \frac{\mu_f \epsilon^2}{512L_f^2 na^2}. \tag{6}$$

Данная оценка требует точности решения по функии на k-ом уровне рекурсии вспомогательной задачи порядка ϵ^{2k}. Таким образом, учитывая максимальную глубину рекурсии, равную $n-1$, получаем, что нам потребуется в худшем случае на каждом шаге алгоритма решать задачу с точностью ϵ^{2n-2} по функции.

Интуитивно понятно, что для функции с липшицевым градиентом "большая" величина ортогональной компоненты при близости к решению не сможет сильно уменьшиться, а следовательно, изменить направление, а с другой стороны, если эта компонента малая, и решалась вспомогательная задача на многомерном параллепипеде Q_k, то мы можем взять эту точку как решение всей задачи на Q_k. Представим стратегию, реализующую эти идеи.

Здесь и далее нам понадобится следующая простая лемма:

Лемма 0.2.

$$\forall a, b \in \mathbb{R}, |a - b| \leq |b| \Rightarrow ab \geq 0$$

Доказательство. Если $b > 0$ и $a \leq b$, тогда условие из леммы эквивалентно следующему:

$$b - a \leq b \Rightarrow a \geq 0.$$

Если $b > 0$ и $a \geq b$, тогда $a \geq 0$.

Случай отрицательного b доказывается аналогично. □

Теорема 0.1. *Функция f выпуклая с липшецевым градиентом с константой L_{xx}. Точка $\boldsymbol{x}_*$ решение вспомогательной задачи оптимизации, $\boldsymbol{x}$ - ее приближение, $\Delta = \|\boldsymbol{x}_* - \boldsymbol{x}\|$ есть расстояние между ними.*

Тогда если приближение удовлетворяет следующему условию:

$$\Delta \leq \frac{|f'_\perp(\boldsymbol{x})|}{L_{xx}},$$

то множество, выбранное на основе градиента в этой точке, содержит решение исходной задачи на квадрате.

Доказательство. Заметим, что множества выбирается правильно, если знак производной в решении вспомогательной задачи по фиксированной переменной совпадает со знаком производной в ее приближении.

Из леммы 0.2 следует, что для того чтобы совпали знаки, достаточно потребовать

$$|f'_\perp(\mathbf{x}_*) - f'_\perp(\mathbf{x})| \leq |f'_\perp(\mathbf{x})|$$

Используя липшецевость градиента получаем утверждение теоремы. □

Описанная в теореме оценка крайне не эффективна, если модуль ортогональной компоненты стремительно убывает при приближении к точке-решению, поэтому сформулируем альтернативное условие остановки:

Теорема 0.2. *Пусть f есть M_f-липшецева выпуклая с L_f-липшецевым градиентов. Пусть решается вспомогательная задача на множестве $\tilde{Q}$ для задачи на множестве Q. Точка $\boldsymbol{x}_*$ есть решение задачи оптимизации на $\tilde{Q}$, $\boldsymbol{x}$ - ее приближение, $\Delta \geq \|\boldsymbol{x}_* - \boldsymbol{x}\|$ - верхняя оценка расстояния между ними.*

Тогда для достижения точности решения ϵ на множестве Q в точке $\boldsymbol{x} \in \tilde{Q} \subset Q$ по функции следующего условия достаточно:

$$\Delta \leq \frac{\epsilon - R|f'_\perp(\boldsymbol{x})|}{L_f + M_f R},$$

где $R = a\sqrt{n}$ - это размер диагонали в исходном гиперкубе.

Доказательство. Из леммы 0.1 мы имеем:

$$g \in \partial f(\mathbf{x}_*) : g_\| = 0.$$

Тогда по определению субградиента:

$$f(\mathbf{x}^*) - f(\mathbf{x}_*) \geq (g, \mathbf{x}^* - \mathbf{x}_*)$$

Используем неравенство Коши-Буняковского-Шварца:

$$f(\mathbf{x}_*) - f(\mathbf{x}^*) \leq -(g, \mathbf{x}^* - \mathbf{x}_*) \leq$$

$$\leq \|g\|\|\mathbf{x}^* - \mathbf{x}_*\| \leq \|g\|a\sqrt{n}$$

С другой стороны, из липшецевости функции мы имеем:

$$f(\mathbf{x}) - f(\mathbf{x}_*) \leq M_f \Delta$$

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq M_f \Delta + \|g\|a\sqrt{n} = M_f \Delta + |f'_\perp(\mathbf{x}_*)|R$$

Из липшецевости градиента:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq M_f \Delta + \left(|f'_\perp(\mathbf{x})| + L_f \Delta\right) R$$

Тогда для достижения точности ϵ по функции в точке $\mathbf{x}$ для исходной задачи достаточно следующего условия:

$$M_f \Delta + \|g\|a\sqrt{n} = M_f \Delta + \left(|f'_\perp(\mathbf{x})| + L_f \Delta\right) R \leq \epsilon$$

$$\Delta \le \frac{\epsilon - R|f'_\perp(\mathbf{x})|}{M_f + L_f R}$$

□

В этой теореме подразумевается, что ϵ - есть точность задачи по функции, с которой нужно решить задачу оптимизации на Q, при условии, что мы находимся на Q_k. Т.е. это ϵ для подзадач, возникающих в подзадачах, отличается от ϵ для исходной задачи на начальном гиперкубе.

Определим нашу адаптивную стратегию. Мы решаем вспомогательную задачу оптимизации, пока не выполнено условие на точность по аргументу:

$$\Delta \le \max\left\{\frac{|f'_\perp(\mathbf{x})|}{L}, \frac{\epsilon - R|f'_\perp(\mathbf{x})|}{M + LR}\right\}. \tag{7}$$

Эту стратегию мы назвали **CurGrad**(Current Gradient). Заметим, что если нам доступен не сам градиент $\nabla f(\mathbf{x})$, а некоторое его приближение $\nu(\mathbf{x})$ при условии

$$\|\nabla f(\mathbf{x}) - \nu(\mathbf{x})\| \le \tilde{\delta},$$

то условие 7 будет выполнено, если

$$\max\left(\frac{1}{L}, \frac{R}{M + LR}\right)\tilde{\delta} + \Delta \le \max\left\{\frac{|\nu_\perp(\mathbf{x})|}{L}, \frac{\epsilon - R|\nu_\perp(\mathbf{x})|}{M + LR}\right\},$$

или после упрощения

$$\frac{1}{L}\tilde{\delta} + \Delta \le \max\left\{\frac{|\nu_\perp(\mathbf{x})|}{L}, \frac{\epsilon - R|\nu_\perp(\mathbf{x})|}{M + LR}\right\}. \tag{8}$$

Таким образом, для каждой задачи оптимизации на многомерном прямоугольном параллепипеде Q каждую вспомогательную задачу мы решаем до тех пор, пока не выполнена оценка 8. Причем, если мы остановили решение вспомогательной задачи в точке, в которой выполнено это неравенство для правой части максимума, то мы останавливаем решение на множестве Q.

Для того, чтобы знаки $f'_\perp(\mathbf{x})$ и $\nu_\perp(\mathbf{x})$ совпдали, достаточно потребовать, чтобы $\tilde{\delta} \le |\nu_\perp(\mathbf{x})|$. Если мы не остановили наш метод после выполнения условия 8, то

$$\frac{|\nu_\perp(\mathbf{x})|}{L} \ge \frac{\epsilon - R|\nu_\perp(\mathbf{x})|}{M + LR}$$

$$|\nu_\perp(\mathbf{x})| \geq \frac{L}{M+2LR}\epsilon.$$

Таким образом, если у нас есть итеративный метод, который генерирует $\nu_k(\mathbf{x}) : \|\nu_k(\mathbf{x}) - \nabla f(\mathbf{x})\| \to_{k\to\infty} 0$, то условие $C\delta \leq |\nu(\mathbf{x})|$ будет выполнено за конечное число итераций. Если же это не выполнено, то мы можем посчитать $\nu(\mathbf{x})$ согласно оценке **??**.

Откуда взять условие $\|\nu_k(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \tilde{\delta}$? Пусть $\nu(\mathbf{x})$ есть δ-субградиент для функции f в точке $\mathbf{x}$. Пользуясь тем, что $f(\mathbf{x}) = S(\mathbf{x}, \mathbf{y}(\mathbf{x}))$ и тем, что для такой задачи $\nabla_x S(\mathbf{x}, \mathbf{y}_\delta)$, где $\mathbf{y}_\delta$ есть решение задачи $\max_{\mathbf{y}} S(\mathbf{x}, \mathbf{y})$ с точностью δ по функции, есть δ-субградиент. При этом в силу липшецевости S.

$$\|\nabla_\delta f(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq L_{yy}\|\mathbf{y}(\mathbf{x}) - \mathbf{y}_\delta\|$$

Таким образом, для того, чтобы правильно определить оставшееся множество, достаточно выполнения неравенства 8 и неравенства

$$L_{yy}\|\mathbf{y}(\mathbf{x}) - \mathbf{y}_\delta\| \leq |\left(\nabla_\delta f(\mathbf{x})\right)_\perp|. \tag{9}$$

Выше обозначенный способ требует сходимости внутренней задачи по аргументу. Однако если $S(\mathbf{x}, \mathbf{y})$ - μ_y-сильно выпукла по $\mathbf{y}$, то эти условия можно заменить на условия сходимости по функции. С другой стороны, если $\rho(\mathbf{x}, \partial Q) \geq \sqrt{\frac{\delta}{2L_f}}$, то мы имеем из **??** следующее неравенство:

$$\|\nabla_\delta f(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \sqrt{\frac{L_f\delta}{2}}. \tag{10}$$

Приведем оценку сходимости для нашего метода.

Теорема 0.3. *Если функция f выпуклая и -липшецева, тогда для достижения ϵ по функции следующего количества итераций достаточно:*

$$N = \left\lceil \log_2 \frac{\sqrt{n}Ma}{\epsilon} \right\rceil \tag{11}$$

где a размер исходного квадрата Q.

References

[1] Gasnikov A. Universal gradient descent // MIPT — 2018, 240 p.

[2] Pasechnyk D.A., Stonyakin F.S. One method for minimization a convex Lipchitz continuous function of two variables on a fixed square // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1812.10300.pdf

[3] Nesterov U.E. Methods of convex optimization // M.MCNMO — 2010, 262 p.

[4] Anaconda[site]. At available: https://www.anaconda.com

[5] Danskin, J.M.: The theory of Max-Min, with applications. J. SIAM Appl. Math.14(4) (1966)

[6] Fedor S. Stonyakin, Mohammad S. Alkousa, Alexander A. Titov,and Victoria V. Piskunova1 On Some Methods for Strongly Convex Optimization Problems with One Functional Constraint // ...

[7] Olivier Devolder Exactness, Inexactness and Stochasticityin First-Order Methods for Large-ScaleConvex Optimization // UCL — 2013,

[8] Need Reference To Book with Inexact Ellipsoids

[9] B.T. Polyak. The Introduction to Optimization // Moscow, Science - 1983

[10] Repository with code: https://github.com/ASEDOS999/Optimization-Halving-The-Square