Contents

1 Описание метода

Рассмторим задачу минимизации функции от двух переменных, заданной на квадрате:

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

где f - это выпуклая функция, Q - квадрат на плоскости. Здесь и далее будем считать, что стороны квадрата сориентированы параллельно осям текущей системы координат. Очевидно, что это не предположение не существенно, поскольку для любого квадрата существует тривиальное афинное преобразование, которое повернет этот квадрат так, чтобы он соответствовал условию

Тогда рассмотрим метод двумерной дихотомии. Его основные этапы следующие:

1. Провести отрезок через центр квадрата параллельно оси Ox.
2. Решить одномерную задачу оптимизации на этом отрезке с некоторой точностью δ
3. Вычислим градиент в полученной точке и отбросим прямоугольник, в который смотрит этот градиент
4. Повторить шаги 2-3 для вертикального отрезка через центр
5. Повторять шаги 1-4, пока не достигнута необходимая точность по функции.

Заметим, что на шаге 3 мы не нуждаемся во всем градиенте, а только в его ортогональной компоненте, т.е. производной по той переменной, которая была фиксировона на текущем отрезке.

Для решения одномерной задачи мы можем использовать любой метод одномерной оптимизации. Однако, в следующих разделах мы покажем, что дифференцируемость функции - существенное условие для сходимости метода, поэтому без ограничения общности мы можем и будем использовать метод дихотомии, использующий производную в центре отрезка, поскольку этот метод гарантирует линейную скорость сходимости и при этом является одноточечным, в отличии, например, от метода золотого сечения или метода дихотомии нулевого порядка.

Утверждается, что для некоторой точности решения вспомогательной задачи δ данный метод будет сходиться к решению по функции. Это будет обсуждено в следующей секции.

Сейчас обсудим корректность алгоритма. Здесь и далее мы будем использовать следующую нотацию. Если у нас определен некоторый сегмент и есть некоторый вектор $\mathbf{g}$, то $\mathbf{g}_{\|}$ - его проекция на этот отрезок, а $\mathbf{g}_{\perp}$ - его перпендикулярная компонента.

Здесь и далее нам понадобятся стандартные понятия субдифференциала и субградиента, которые можно найти, например, в [3].

Нам понадобится следующая лемма.

Лемма 1.1. *Если $\boldsymbol{x}_*$ решение одномерной задачи оптимизации:*

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

Доказательство. Если $\mathbf{x}_*$ есть внутренняя точка, то производная по нефиксированной переменной равна нулю в силу того, что $\mathbf{x}_*$ - минимум. Тогда с учетом того, что $\nabla f(\mathbf{x}_*) \in \partial f(\mathbf{x}_*)$, получаем утверждение из теоремы.

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

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

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

Далее предполагаем, что мы решаем задачу оптимизации на горизонтальном сегменте и $\mathbf{x}_*$ есть правый конец этого отрезка. Доказательство на три других случая (другой конец горизонтального сегмента и два конца вертикального сегмента). В нашем случае:

$$\partial f(\mathbf{x}_*) = \{\nabla f(\mathbf{x}_*)\},$$
$$N(\mathbf{x}|Q) = \{\mathbf{a}|\mathbf{a}_{\perp} = 0, \mathbf{a} \parallel \geq 0\}$$

Заметим, что $\left(\nabla f(\mathbf{x}_*)\right)_{\|} \leq 0$, потому что, если бы это было не так, то очевидно $\mathbf{x}_*$ - не минимум функции f на этом сегменте. Тогда выбрав $\mathbf{a}$, такой что $\mathbf{a}_{\|} = -\left(\nabla f(\mathbf{x}_*)\right)_{\|}$, получаем субградиент из условия:

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

□

Теорема 1.1. *Если функция f выпуклая непрерывно дифференцируемая функция, то для решения любой одномерной задачи существует некоторая*

его окрестность, такая что, если выбрать прямоугольник на основе перпендикулярной компоненты градиента в любой точке этой окрестности, выбор будет такой же, как и в случае использования градиента в точке-решении.

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

2 Одномерная задача

В этой секции мы опишем достаточную точность решения вспомогательной задачи на отрезке. Мы будем использовать следующие обозначения:

$$(x_*, y_*) \text{ or } \mathbf{x}_* \text{ – solution of one-dimensional problem}$$
$$\delta = |x - x_*| \text{ – distance}$$

В работе [2] была доказана следующая теорема:

Теорема 2.1. *Пусть функция f выпуклая и удовлетворяет условие Липшица с константой M, а ее градиент - с константой L. Тогда если каждая одномерная задача решена со следующей точностью по аргументу*

$$\delta \leq \frac{\epsilon}{2La(\sqrt{2}+\sqrt{5})(1-\frac{\epsilon}{Ma\sqrt{2}})} \tag{1}$$

то метод двумерной дихотомии сходится к решению с точностью ϵ по функции.

Данная стратегия всегда требует решать задачу с точностью порядка ϵ. А также данный метод может не гарантировать сходимости по аргументу. Такой пример также был рассмотрен в работе [2]. Далее мы будем называть эту стратегию **ConstEst**(Constant Estimate).

Теперь разработаем стратегию, которая гарантирует сходимость по аргументу. Заметим, что прямоугольник выбирается правильно, если знак производной в решении вспомогательной задачи совпадает со знаком производной в ее приближении:

$$f'_\perp(\mathbf{x}_*) f'_\perp(\mathbf{x}_* + \delta) > 0 \tag{2}$$

Здесь и далее мы будем использовать следующую простую лемму:

Лемма 2.1. $\forall a, b \in \mathbb{R}\backslash\{0\}, |a - b| \leq |b| \Rightarrow \operatorname{sign} a = \operatorname{sign} b$

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 2.2. *Пусть функция f выпуклая с липшецевым градиентом с константой L. Точка $\boldsymbol{x}_*$ решение одномерной задачи оптимизации, $\boldsymbol{x}$ - ее приближение.*

Тогда если приближение удовлетворяет следующему условию:

$$\delta \leq \frac{|f'_\perp(\boldsymbol{x})|}{L},$$

то прямоугольник, выбранный на основе градиента в этой точке, содержит решение исходной задачи на квадрате.

Доказательство. Из леммы 2.1 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 перпендикулярной компоненты стремительно убывает при приближении к точке-решению, поэтому сформулируем альтернативное условие остановки:

Теорема 2.3. *Пусть f есть M-липшецева выпуклая с L-липшецевым градиентов. Точка $\boldsymbol{x}_*$ есть решение одномерной задачи оптимизации, $\boldsymbol{x}$ - ее приближение, $\delta = \|\boldsymbol{x}_* - \boldsymbol{x}\|$ - верхняя оценка расстояния между ними.*

Тогда для достижения точности ϵ в точке $\boldsymbol{x}$ следующего условия достаточно:

$$\delta \leq \frac{\epsilon - LR|f'_\perp(x)|}{L + MR},$$

where $R = a\sqrt{2}$ is size of current square.

Доказательство. Из леммы 1.1 мы имеем:

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

Then the following inequallity is true by diffenition of subgradient:
Тогда по определению субградиента:

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

Используем неравенство Коши-Буняковского-Шварца:

$$f(\mathbf{x}_*) - f(\mathbf{x}^*) \leq -(g, \mathbf{x}^* - \mathbf{x}_*) \leq$$
$$\leq \|g\|\|\mathbf{x}^* - \mathbf{x}_*\| \leq \|g\| a\sqrt{2}$$

С другой стороны, из липшецевости функции мы имеем:

$$f(\mathbf{x}) - f(\mathbf{x}_*) \leq M\delta$$

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq M\delta + \|g\| a\sqrt{2} = M\delta + |f'_{\perp}(\mathbf{x}_*)|R$$

Из липшецевости градиента:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq M\delta + (|f'_{\perp}(\mathbf{x})| + L\delta)\, R$$

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

$$M\delta + \|g\| a\sqrt{2} = M\delta + (|f'_{\perp}(\mathbf{x})| + L\delta)\, R \leq \epsilon$$

$$\delta \leq \frac{\epsilon - |f'_{\perp}(\mathbf{x})|}{M + LR}$$

□

Then our addaptive strategy is following. One is to calculate untill the following condition is met:
Определим нашу адаптивную стратегию. Мы решаем задачу одномерной оптимизации, пока не выполнено условие на точность по аргументу:

$$\delta \leq \max\left\{\frac{|f'_{\perp}(\mathbf{x})|}{L}, \frac{\epsilon - |f'_{\perp}(\mathbf{x})|}{M + LR}\right\}. \tag{3}$$

Причем, если выполнено условие из теоремы 2.3, мы останавливаем весь метод. Эту стратегию мы назвали **CurGrad**(Current Gradient).

3 Сходимость

В этой сецкии будут приведены оценки для количества итераций глобального метода для достижения точности по функции. В данном разделе под итерацией подразумевается одна итерация метода двумерной дихотомии, в результате которой квадрат уменьшается вдвое.

Теорема 3.1. *Если функция f выпуклая и -липшецева, тогда для достижения ϵ по функции следующего количества итераций достаточно:*

$$N = \left\lceil \log_2 \frac{\sqrt{2}Ma}{\epsilon} \right\rceil \tag{4}$$

где a размер исходного квадрата Q.

Доказательство для стратегии **ConstEst** было проведено в работе [**?**]. Для стратегии **CurGrad** данная оценка есть очевидное следствие липшецевости функции и сходимости по аргументу.

Однако мы можем улучшить нашу оценку, если учтем сходимость по аргументу.

Теорема 3.2. *Пусть f - выпуклая функция.*

Если

1. Функция f имеет L-липшецев градиент

2. $\exists \boldsymbol{x}^* \in Q : \nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$

3. Стратегия для решения одномерной задачи обеспечивает сходимость по аргументу

тогда для достижения точности ϵ по функции следующего количества итераций достаточно:

$$N = \left\lceil \frac{1}{2} \log_2 \frac{La^2}{4\epsilon} \right\rceil, \tag{5}$$

где a - размер исходного квадрата Q.

Доказательство. Для всех выпуклых дифференцируемых функций следующее неравенство верно (доказательство можно найти в [3]):

$$f(\mathbf{x}) - f(\mathbf{x}^*) - (f'(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$$

По условию теоремы существует такая точка $\mathbf{x}^*$, что $f'(\mathbf{x}^*) = 0$. Заметим, что эта точка является решением задачи. Тогда наше неравенство примет вид:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$$

После N итераций имеем следующую оценку:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq L\left(\frac{a}{2^N}\right)^2,$$

что доказывает оценку из теоремы. □

4 Двойственные задачи

Данный метод предполагает особый интерес в приложении решения двойственных задач для задач с двумя ограничениями. А именно, для решения задач вида:

$$\phi(\lambda_1, \lambda_2) \to \min_{\lambda \geq 0}, \tag{6}$$

$$\text{where } \phi = -\min_{\mathbf{x}} \left(f(\mathbf{x}) + \lambda_1 g_1(\mathbf{x}) + \lambda_2 g_2(\mathbf{x})\right) \tag{7}$$

$$\mathbf{x}(\lambda) = \arg\min_{\mathbf{x}} \Phi(\mathbf{x}, \lambda)$$

In this section we will discuss how transform this task to task of the task of minimization on square, what is derivative and lipschitz constants. Also, there is description of way for to calculate the value of function ϕ and its derivative.

В данной секции мы обсудим как свести это к нашей задаче, как вычислить липшецевы константы и каким образом вычислять $\mathbf{x}(\lambda)$.

Для начала приведем задачу к задаче минимизации на квадрате. Согласно [1] (see ex. 4.1), мы имеем следующую локализацию на решение:

$$\|\lambda\|_1 \leq a = \frac{1}{\gamma}\left(f(\overline{\mathbf{x}}) - \min_{\mathbf{x}} f(\mathbf{x})\right), \tag{8}$$

$$\text{where } \overline{\mathbf{x}} : g_i(\overline{\mathbf{x}}) < 0, \gamma = \min_i[-g_i(\overline{\mathbf{x}})] \tag{9}$$

Согласно этому утверждению, мы можем локализовать λ^* в квадрате $Q = [0, a]^2$. В таком случае мы свели нашу задачу оптимизации к следующему виду:

$$\phi(\lambda_1, \lambda_2) \to \min_{\lambda \in Q} \tag{10}$$

Градиент функции ϕ мы будем считать согласно хорошо известной теорема Демьянова-Данскина-Рубинова, см. [5].

Теорема 4.1. *Пусть* $\phi(\lambda) = \min\limits_{x \in X} \Phi(x, \lambda)$ *для всех* $\lambda \geq 0$, *где* Φ *это гладкая выпуклая функция по* λ. *Тогда*

$$\nabla \phi(\lambda) = F'_\lambda \left(x(\lambda), \lambda\right)$$

Условия теоремы выполнено в нашем случае и мы получаем:

$$\phi'_{\lambda_k}(\lambda) = g_k \left(\mathbf{x}(\lambda)\right) \tag{11}$$

Кроме этого, мы нуждаемся в константе Липшица для градиента. В работе [6] сформулирована и доказана следующая теорема:

Теорема 4.2. *Let* $f(x)$ *be a* μ_f*-strongly convex function, the function* $g(x)$ *satisfies the Lipschitz condition with a constant* M_g. *Then the function* $\phi(\lambda) = \min_{\boldsymbol{x}} \left(f(\boldsymbol{x} + \lambda_1 g_1(\boldsymbol{x}) + \lambda_2 g_2(\boldsymbol{x})\right)$ *defined in 17, where* $x(\lambda) = \arg\min\limits_x (f(x) + \lambda g(x))$, *has Lipschitz smooth gradient with constant* $L_{\phi'} = \frac{M_g^2}{\mu_f}$

В случае размерности 2 по λ доказательство повторяется практически в точности. В таком случае, $g(\mathbf{x})$ есть вектор-функция.

Основная сложность решения седловых задач есть то, что вычисление $\mathbf{x}(\lambda)$ точно в большинстве случаев невозможно. Следовательно, мы не имеем доступа к точному значению градиента. Это приводит к следующим проблемам:

1. Шаг дихотомии на отрезке
2. Проверка стоп-условия из стратегии **CurGrad**
3. Выбор прямоугольника

Заметим, что нас в каждой из задач интересуют следующие величины:

1. Проекция градиента на отрезок, т.е. $g_\parallel(\mathbf{x}(\lambda))$.
2. Значение разниц $\delta - \frac{|g_\perp(\mathbf{x}(\lambda))|}{L}$ и $\delta - \ldots$, где L это липшецева константа для градиента ϕ.

3. Перпендикулярная компонента, т.е. $g_2(\mathbf{x}(\lambda))$.

Согласно [2], мы можем вычислять производные неточно, для того чтобы выбрать прямоугольник с ϵ-решением. Так, если δ - это точность по аргументу текущего приближения одномерной задачи, и Δ - точность вычисления градиента в точке по величине градиента, то следующего условия достаточно для нужного выбора прямоугольника:

$$2\Delta + L\delta \leq \frac{\epsilon}{2a(\sqrt{2}+\sqrt{5})},$$

где L - это липшецева константа градиента для двойственной задачи. Мы можем увидеть, что задача выбора сегмента с решением одномерной задачи является одномерным вариантом той же проблемы. Это приводит нас к следующей теореме.

Теорема 4.3. *Если одномерная задача решается со следующей точностью* $delta = \frac{\epsilon}{4La(\sqrt{2}+\sqrt{5})}$ *по аргументо, тогда, для того чтобы выбрать прямоугольник с* ϵ*-решением, достаточно вычислять* $\boldsymbol{x}(\lambda)$ *в текущей точке со следующей точностью:*

$$\|\boldsymbol{x} - \boldsymbol{x}(\lambda)\| \leq \frac{1}{M_g} \frac{\epsilon}{8a(\sqrt{2}+\sqrt{5})} \tag{12}$$

Для того, чтобы выбрать сегмент с решением одномерной задачи, достаточно вычислять $\boldsymbol{x}(\lambda)$ *в центре текущего сегмента со следующей точностью:*

$$\|\boldsymbol{x} - \boldsymbol{x}(\lambda)\| \leq \frac{1}{M_g} \frac{\epsilon}{4a(\sqrt{2}+\sqrt{5})} \tag{13}$$

С другой стороны, для каждого случая интересен только знак соответствующих выражений. Для этого мы будем применять лемму 2.1. Согласно этой лемме, мы имеем следующие достаточные условия для остановки вычисления $\mathbf{x}(\lambda)$ для выше обозначенных случаев:

1. $|g_{\|}(\mathbf{x}) - g_{\|}(\mathbf{x})| \leq |g_{\|}(\mathbf{x})|$
2. $\frac{1}{L}\Big||g_{\perp}(\mathbf{x})| - |g_{\perp}(\mathbf{x}(\lambda))|\Big| \leq \Big|\delta - \frac{|g_{\perp}(\mathbf{x})|}{L}\Big|$
3. $|g_{\perp}(\mathbf{x}) - g_{\perp}(\mathbf{x}(\lambda))| \leq |g_{\perp}(\mathbf{x})|$

Тогда сформулируем утверждения:

Теорема 4.4. *Пусть* g_k *есть* L_{g_k}*-липшецевы функции. Тогда следующие утверждения верны.*

1. For to make the dichotomy step correctly one can calculate $\boldsymbol{x}(\lambda)$ *untill the following condition is approached:*

1. Для того, чтобы сделать шаг дихотомии корректно, т.е. сохраняя сходимость по аргументу, достаточно вычислять $\boldsymbol{x}(\lambda)$*, пока не выполнено следующее условие*

$$L_{g_\parallel}\|\boldsymbol{x} - \boldsymbol{x}(\lambda)\| \le |g_\parallel(\boldsymbol{x})|.$$

2. For to test stop condition correctly one can calculate $\boldsymbol{x}(\lambda)$ *untill the following condition is approached:*

$$\frac{L_{g_\perp}}{L}\|\boldsymbol{x} - \boldsymbol{x}(\lambda)\| \le \left|\delta - \frac{|g_\perp(\boldsymbol{x})|}{L}\right|,$$

where δ *is a distance between* λ *and solution of one dimensional task* λ_**.*

3. For to select rectangle according to strategy one can calculate $\boldsymbol{x}(\lambda)$ *untill the following condition is approached:*

$$L_{g_\perp}\|\boldsymbol{x} - \boldsymbol{x}(\lambda)\| \le |g_\perp(\boldsymbol{x})|.$$

There are two interesting remarks.

Firstly, the calculating of $\mathbf{x}(\lambda)$ does not depend on required accuracy ϵ for function ϕ. This result is unique for our method and one can not see such effect in other inexact methods (see the next section).

Secondly, this strategies from written above theorem does not guarantee that the calculating of $\mathbf{x}(\lambda)$ will stop. For example, let's consider the first condition. If $g_1(\mathbf{x}(\lambda)) = 0$ and $|g_1(\mathbf{x})|$ decreases faster than $L_{g_1}\|\mathbf{x} - \mathbf{x}(\lambda)\|$ with decreasing of $\|\mathbf{x} - \mathbf{x}(\lambda)\|$ than the stop condition will not be approached. This case is looked as extra specific and does not observed in our experiments but it can take place.

5 Complexity

Let's estimate complexity of one-dimensional problem.

For the strategy **ConstEst** we have that each one-dimensional problem needs exactly the following iterations of one-dimensional method:

$$\log_2 \frac{2La(\sqrt{2}+\sqrt{5})(1 - \frac{\epsilon}{Ma\sqrt{2}})}{\epsilon} = O\left(\log\frac{1}{\epsilon}\right)$$

Now let's condider the strategy **CurGrad**. Let the modul of the perpendicular component at point-solution $|f'_\perp(\mathbf{x}_*)|$ is equal to $\tilde{\epsilon}$. A point from segment $\mathbf{x}$ is its approximation and δ is a distance between them.

If we use the dichotomy method and N is a number of current dichotomy iteration we have from L-Lipschitz continious gradient the following estimates for derivative at $\mathbf{x}$:

$$\tilde{\epsilon} - La2^{-N} \leq |f'_\perp(\mathbf{x})| \leq \tilde{\epsilon} + La2^{-N}$$

Accordint to this estimates for to approach a the first alternative from adaptive estimate 3 the following condition is sufficient:

$$a2^{-N} \leq \frac{\tilde{\epsilon} - La2^{-N}}{L}$$

Similarly for the second alternative:

$$a2^{-N} \leq \frac{\epsilon - \tilde{\epsilon} - La2^{-N}}{M + LR} \leq \frac{\epsilon - \tilde{\epsilon}}{M + LR} - La2^{-N}$$

Accordint to this we have that for to approach estimate 3 we need the following iterations number:

$$1 + \log_2 \min\left\{\frac{La}{\tilde{\epsilon}}, \frac{(M + LR)a}{\epsilon - \tilde{\epsilon}}\right\}$$

Of course, we assume in this estimate that $\tilde{\epsilon}$ and $\epsilon - \tilde{\epsilon}$ are positive. If it is not true we can trash a bad alternative.

The bad case of this iterations number's estimate occurs when

$$\tilde{\epsilon} = \frac{L}{M + L(R+1)}\epsilon.$$

In the bad case we have the following estimate for iterations number:

$$1 + \log_2 \frac{(M + L(R+1))a}{\epsilon} = O\left(\log \frac{1}{\epsilon}\right)$$

From this estimate we see that the our adaptive strategy needs the same iterations number in the bad case as the strategy **ConstGrad** needs always.

But for complexity estimation we have that for to solve one-dimensional problem we needs $O\left(\log \frac{1}{\epsilon}\right)$ iterations.

Теорема 5.1. *For to approach accuracy ϵ on function our method needs the following number of calculating function f and its derivatives of the first order:*

$$O\left(\log^2 \frac{1}{\epsilon}\right)$$

Доказательство. We have that for to solve one-dimensional problem one needs not more than

$$\log_2 \frac{Ca}{\epsilon}$$

iteration, where C is a constant determined by the used strategy and function's parameter and a is a size of current segment.

On each iteration of our method we solve to problems on segments of size $a_N = a2^{-N}$. Let $N_{\max 1} = \left\lfloor \log_2 \frac{Ca}{\epsilon} \right\rfloor$ is a maximal number of global method which the estimation for iteration number of one-dimensional problem is positive. And $N_{\max 2} = \left\lceil \log_2 \frac{La\sqrt{2}}{\epsilon} \right\rceil$ is estimation from 4.

$$N_{\max} := \min\{N_{\max 1}, N_{\max 2}\} = \log_2 \frac{1}{\epsilon} + O(1)$$

we have the following number of derivatives and function values calculation:

$$\sum_{k=0}^{N_{\max}} \log_2 \frac{Ca2^{-k}}{\epsilon} =$$

$$= \sum_{k=0}^{N_{\max}} \log_2 \frac{Ca}{\epsilon} - \sum_{k=0}^{N_{\max}} k =$$

$$= N_{\max} \log_2 \frac{Ca}{\epsilon} - \frac{1}{2} N_{\max}^2 + O\left(\log \frac{1}{\epsilon}\right) =$$

$$= \frac{1}{2} \log_2^2 \frac{1}{\epsilon} + O\left(\log \frac{1}{\epsilon}\right) =$$

$$= O\left(\log^2 \frac{1}{\epsilon}\right)$$

□

In the case of dual problems the complexity of function's parameters calculation depends on ϵ too. But accordinп to 4.3 in this case the complexity will change by obvious way:

Теорема 5.2. *For to approach accuracy ϵ on function our method needs the following number of calculation of functions from primal problem and their derivatives:*

$$O(\log^3 \frac{1}{\epsilon})$$

6 Other Inexact Methods

Our optimization method can solve the task of minimization function f on square when the function and its gradient can not be calculated accurately but there are other optimization method for such tasks. In each section one describes some of them and below there is experimental comparison of them with our method.

The fist method is Primal Gradient Method (PGM) with (δ, L, μ) oracle. There are proves in the [7] that this method converges to the solution with accuracy δ:

$$\min_k f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{LR^2}{2} \exp\left(-k\frac{\mu}{L}\right) + \delta,$$

where $R = \|\mathbf{x}_0 - \mathbf{x}^*\|$ in our task. Moreover, in the [7] it is proved that for the function

$$f(\mathbf{x}) = \min_{\mathbf{u}} \left(\Psi(\mathbf{x}, \mathbf{u}) + \mathbf{u}^\top A\mathbf{x}\right)$$

there is following (δ, L, μ) oracle:

$$f_{\delta,L,\mu}(\mathbf{x}) = \Psi(\mathbf{x}, \mathbf{u}_\mathbf{x}) - \xi$$

$$g_{\delta,L,\mu}(\mathbf{x}) = A\mathbf{u}_\mathbf{x}$$

with parameters $\delta = 3\xi$, $L = \frac{2\lambda_{\max}(A^\top A)}{\mu(G)}$, $\mu = \frac{\lambda_{\min}(A^\top A)}{2L(G)}$ if $\mathbf{u}_\mathbf{x}$ is a solution approximation of $\mathbf{u}^*$ for current $\mathbf{x}$ with accuracy ξ on function.

The second method is Fast Gradient Method with (δ, L, μ) oracle. this method converges to the solution with accuracy δ:

$$\min_k f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \min\left(\frac{4LR^2}{k^2}, LR^2 \exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right)\right) + C_k\delta,$$

where $C_k = \min\left(\frac{k}{3} + \frac{12}{5}, 1 + \sqrt{\frac{L}{\mu}}\right)$. The method's description and proves for it is in the [7] too. This method has significantly better convergence rate for ill-conditioned problems.

The third having to be discussed method is inexact ellipsoid method. The ellipsoid method with ϵ-subgradient instead usual subgradient converges to a solution with accuracy ϵ:

$$\min_k f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \max_{\mathbf{x} \in Q} |f(\mathbf{x})| \exp\left(-\frac{k}{8}\right) + \delta$$

It is proved in [8]. Moreover, in [9] there is proved that for the function

$$f(\mathbf{x}) = \min_{\mathbf{u}} \Psi(\mathbf{x}, \mathbf{u})$$

the following statement is met:

$$\Psi(\mathbf{x}, \mathbf{u}_{\mathbf{x},\epsilon}) \in \partial_\epsilon f(\mathbf{x}),$$

if $\mathbf{u}_{\mathbf{x},\epsilon}$ is such point that $\Psi(\mathbf{x}, \mathbf{u}_{\mathbf{x},\epsilon}) - \min_{\mathbf{u}} \Psi(\mathbf{x}, \mathbf{u}) \leq \epsilon$.

Note that our method converge by the following way:

$$\min_k f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq Cexp(-k \ln 2),$$

and it is the best theoretical convergence rate from all methods in this section.

7 Tests

In this section we show estimate on number iterations of practice. Also there is comparison work time of our new method with work time of other optimization methods such as inexact gradient methods and inexact ellipsoids method[1]. All code was made in Anaconda 5.3.1 Python 3.6 (see cite [4])

7.1 Comparison With Other Methods

Let's compare our method with inexact ellipsoid method and gradient methods (PGM and FGM) with (δ, L, μ) oracle (see previous subsection 6) on a dual task.

For to find $\mathbf{x}(\lambda)$ we will use gradient descent. There are theoretical result that can help to manange distance $\|\mathbf{x}_k - \mathbf{x}^*\|$ from optimal point to its current approximation. In particular, there are following results:

[1] You can find all code in the repository [10]

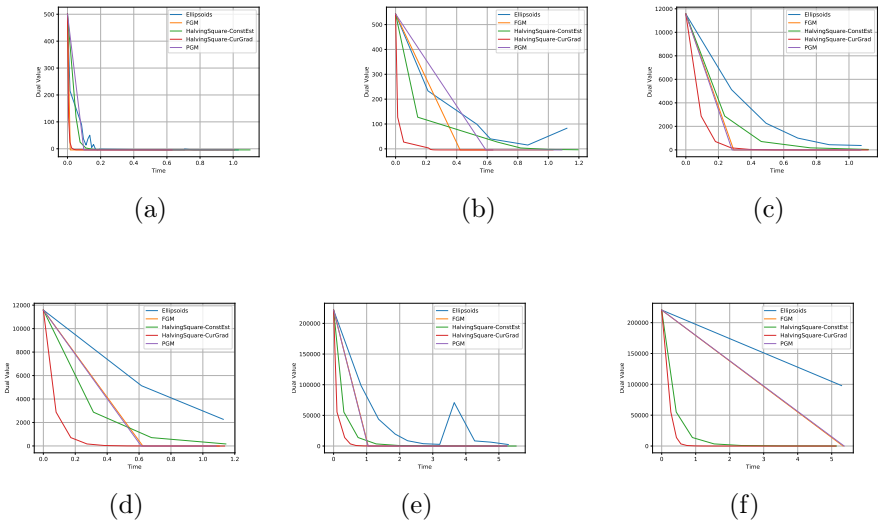

Figure 1: Comparison of different on inexact methods for task with different dimension N and for different required accuracy ϵ: (a) $N = 100, \epsilon = 10^{-3}$; (b) $N = 100, \epsilon = 10^{-10}$; (c) $N = 1000, \epsilon = 10^{-3}$; (d) $N = 1000, \epsilon = 10^{-10}$; (e) $N = 10000, \epsilon = 10^{-3}$; (f) $N = 10000, \epsilon = 10^{-10}$.

- If f is a convex function with L-Lipschitz continious gradient then gradient descent with step $\alpha_k = \frac{1}{L}$ converges with speed

$$\|f(\mathbf{x}_k) - f(\mathbf{x}^*)\| \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|}{k+4}$$

- If f is a μ-strong convex function with L-Lipschitz continious gradient then gradient descent with step $\alpha_k = \frac{1}{L+\mu}$ converges with speed

$$\|f(\mathbf{x}_k) - f(\mathbf{x}^*)\| \leq \left(\frac{M-1}{M+1}\right)^k L\|\mathbf{x}_0 - \mathbf{x}^*\|,$$

  where $M = \frac{L}{\mu}$.

The proves for this statements one can find in many books of optimization, for example, in the book [9].

We will use functions where μ is small enough. Therefore, our method for calculating $\mathbf{x}(\lambda)$ will converge to solution according to the first estimate.

For all inexact method we will calculate $\mathbf{x}(\lambda)$ with such accuracy as the method will converge to the solution with same for all methods accuracy ϵ. For PGM, FGM and ellipsoids methods we will calculate $\mathbf{x}(\lambda)$ with accuracy $\frac{\epsilon}{2}$ on function. For our method with the both strategies we will calculate $\mathbf{x}(\lambda)$ untill the conditions from the 10 is approached.

We consider the following prime task:

$$f(\mathbf{x}) = \ln\left(1 + \sum_{k=1}^{n} e^{\alpha x_k}\right) + \beta\|\mathbf{x}\|_2^2 \to \min_{\mathbf{x}\in\mathbb{R}^N} \tag{14}$$

$$g_k(\mathbf{x}) = \langle \mathbf{b}_k, \mathbf{x}\rangle + c_k \leq 0, k = \overline{1, m} \tag{15}$$

$$\tag{16}$$

It is task of minimization the LogSumExp-function with $l2$-regularization. The regularization parameter β determines strong convexity of our task and in the tests one takes $beta = 0.1$. The N is diminsionality of primal task and is determined for different tests below. The parameter α is equal to 1. The parameters c_k are equal to 1 too. The vectors $\{b_k\}_{k=1}^m$ are generated randomly for the each test. The m is equal to dimensionality of dual task and in the current case is equal to 2.

The LogSumExp-problem is L-Lipschitz continious function with M-Lipschitz continious gradient where $L = 1$ and $M = \alpha$. Therefore:

$$L_f = \alpha + 2\beta R, M_f = \alpha^2 + 2\beta,$$

$$\mu_f = 2\beta,$$

where $R = \|\mathbf{x}_0 - \mathbf{x}^*\|$ is the size of initial approximation. The functions g_k are L_k-Lipschitz continious where $L_k = \|b_k\|$.

We introduce the following notation:

$$\phi(\lambda_1, \lambda_2) = -\min_{\mathbf{x}\in\mathbb{R}^N} \left(f(\mathbf{x}) + \lambda_1 g_1(\mathbf{x}) + \lambda_2 g_2(\mathbf{x})\right) \tag{17}$$

In such notations the dual task for the task 14 looks like:

$$\phi(\lambda_1, \lambda_2) \to \min_{\lambda_1,\lambda_2} \tag{18}$$

$$\text{s.t}\, \lambda_1, \lambda_2 \geq 0 \tag{19}$$

Obviously, $\min\limits_{\mathbf{x}} f(\mathbf{x}) \geq 0$. Therefore, according to 8 we can add following conditions on the dual variables:

$$|\lambda_k| \leq \lambda_{\max} = \frac{f(\overline{\mathbf{x}})}{\gamma}, k = 1, 2$$

And we have following task:

$$\phi(\lambda_1, \lambda_2) \to \min_{0 \leq \lambda_k \leq \lambda_{max}}$$

Calculating of function and derivatave value for such task was discussed in the section 4.

We can see on 1 the following results. Firstly, the halving square method with provided in this work strategy **CurGrad** are the fastest method in the most tests tests. This method can be slower than other inexact methods if dimensional of primal task is small or ϵ is big. In particular, this strategy is faster than strategy with constant estimate provided in [2]. It proves that provided by Nesterov method with strategy through gradient is the best method for to solve two dimensional dual task of minimization. Secondly, the gain of this strategy in comparison with other method is increase when the required ϵ decrease. This fact demonstrated important advantange of this strategy: it does not depend on required accuracy strongly. So, this method with constant estimate has strong dependity on it because there is this accuracy in the constant estimate, PGM and inexact ellipsoid method require that the $x(\lambda)$ is found with accuracy depended on ϵ. But halving square with ellipsoid method has not such dependety.

8 Conclusion

We discussed and proved that this method converges to the solution for smooth convex functon. Moreover, in the [2] there is conterexample when the problem is non-smooth and we can not to converge to the solution with the accuracy on function better than a constant.

After it we discussed different strategy for one-dimensional task. Two strategies were considered. The both suggest to use stop conditions that are met when the current approximation is "very near" to the segment's solution. The first compares the distance between them with derivative value in accurate segment's solution but the second compares it with derivative

value in approximation. In the experiment the first has a little better result but it can not be used for real task. The second strategy using derivative value in current approximation is significantly better then constant estimate and does not depend on required accuracy.

But all this strategy are good when the derivative value is high enough. But when the segment is near to the global solution this value will be small. Therefore one needs to make a lot of iterations on segment. For to avoid it we consider additional stop condition for global task on square when in current approximation the derivative value is near to zero.

The most steps of methods assumped that derivative can be calculated accurately. But the main method purpose is to solve dual problems and for it one can not usually calculate it so. That's why we consider different modifications of this method for to solve such problems. The important moment is we don't add some dependence on initial required accuracy in the modificated method.

Finally, we compared our method with new strategy for dual problem to prime LogSumExp problem with two linear constaints with our method with strategy using the constant estimate, primal gradient method and fast gradient method with (δ, L, μ) -oracle and with inexact ellipsoids methods. The Halving Square Method is the fastest of them for enough high dimension (more 100) and for enough high required solution ($1e-3$ and more).

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