Методы зеркальных отображений в задачах выпуклой оптимизации

Ефремов С.В., Якушев А.Ю.

8 июня 2019 г.

1 Введение

Метод зеркального спуска является естественным обобщением метода проекции субградиента в случае обобщения l_2 нормы на более общий случай какой-то функции расстояния. В этой работе мы представляем новый подход к построению субградиентной схемы для различных негладких задач выпуклой структуры. Так мы не ограничиваемся построением стандартного метода зеркального спуска, но рассматриваем его вариацию — метод двойственных усреднений, ускоряем алгоритмы методом подобных треугольников, попутно изучая вопрос выбора констант для его оптимальной работы.

2 Постановка задачи

Рассмотрим следующую задачу выпуклой оптимизации – задана функция $f(x)$ и оракул также выдает ее градиент $\nabla f(x)$. Задача найти $\min_{x \in \mathbb{R}} f(x)$.

2.1 Сопряженная норма:

Определение: Сопряженной нормой $\|\cdot\|_*$ к данной $\|\cdot\|$ называется:

$$\|y\|_* = \max\{\langle y,x\rangle : \|x\| = 1\}$$

Пример: $(\|\cdot\|_p)_* = \|\cdot\|_q, \qquad \frac{1}{p} + \frac{1}{q} = 1$
Неравенство Гельдера:

$$\sum_{k=1}^{n} |x_k\, y_k| \le \left(\sum_{k=1}^{n} |x_k|^p\right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} |y_k|^q\right)^{\frac{1}{q}} \text{ for all } x, y \in \mathbb{C}^n$$

Свойства:

- Двойственная норма $\|\cdot\|_*$ является нормой
- l_2 норма сопряжена сама себе
- Двойственная норма к двойственной норме - исходная норма
- $(\|\cdot\|_1)_* = \|\cdot\|_\infty,\ (\|\cdot\|_\infty)_* = \|\cdot\|_1$
- Обобщенное неравенство Коши Шварца: $\langle y{,}x\rangle \leqslant \|y\|_*\|x\|$, следствие: $\|x\|^2 \pm 2\langle y{,}x\rangle + \|y\|_*^2 \geqslant 0$

2.2 Дивергенция (расстояние) Брэгмана

Попробуем интуитивно ввести понятие обобщенного расстояния, именуемого расстоянием Брэгмана. Для каждой точки y она возвращает расстояние этой точки до x - $V_x(y)$. В самом простом случае можно взять $V_x(y) = \frac{1}{2}\|x-y\|^2,\ \ \nabla V_x(y) = y - x$. Рассмотрим уже классическую запись:

$$\begin{gathered}
\|x_{k+1} - y\|^2 = \|x_{k+1} - x_k\|^2 + \|x_k - y\|^2 - 2\langle x_k - x_{k+1}{,}x_k - y\rangle \\
V_{x_{k+1}}(y) = V_{x_{k+1}}(x_k) + V_{x_k}(y) - \langle \nabla V_{x_{k+1}(x_k)}, x_k - y\rangle
\end{gathered}$$

Для вводимого обобщенного расстояния будем требовать выполнения, кроме того (как будет видно при получении оценок), приятным свойством было бы еще следующее требование:

$$V_x(y) \geqslant \frac{1}{2}\|x-y\|^2$$

Определение: Дивергенцией (расстоянием) Брэгмана называется функция следующая $V_x(y)$. Пусть $S \subseteq \mathbb{R}^n$ - замкнутое выпуклое множество, тогда функция $\phi : S \to \mathbb{R}$ называется прокс-функцией (distance generating function), если ϕ является 1 - сильно выпуклой, т.е.:

$$\phi(y) \geqslant \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{1}{2}\|y - x\|^2, \qquad \forall x{,}y \in S$$

Тогда прокс-функцией индуцируется **расстояние Брэгмана**:

$$V_x(y) = \phi(y) - \phi(x) - \langle \nabla\phi(x), y - x\rangle$$

Заметим, что определение сильной выпуклости зависит от выбора прямой нормы $\|\cdot\|$. Это важное замечание, посольку именно это свойство позволит в будущем подстраивать расстояние под геометрию пространства.

Свойства:

- Аксиома тождества $V_x(x) = 0$
- Совместимость с Евклидовой нормой: $V_x(y) \geqslant \frac{1}{2}\|x - y\|^2 \geqslant 0$
- (Не)равенство треугольника: $\langle -\nabla V_x(y), y - z\rangle = V_x(z) - V_y(z) - V_x(y)$

2.3 Вернемся к задаче

Пусть задано выпуклое замкнутое множество $S \in \mathbb{R}^n$, кроме того, есть алгоритм оптимизации, возвращающий последовательность точек $x_1, \ldots, x_k, \ldots$. Тогда запишем (не)равенство треугольника для расстояния Брэгмана, полагая $y = x_{k+1}, x = x_k$ и произвольный $z \in S$ (который мы в дальнейшем для целостности изложения будем обозначать y)

$$\begin{aligned}
\langle -\nabla V_{x_k}(x_{k+1}), x_{k+1} - z\rangle &= V_{x_k}(z) - V_{x_{k+1}}(z) - V_{x_k}(x_{k+1}) \\
\langle -\nabla V_{x_k}(x_{k+1}), x_{k+1} - y\rangle &= V_{x_k}(y) - V_{x_{k+1}}(y) - V_{x_k}(x_{k+1}) \qquad \text{(baseMD)}
\end{aligned}$$

Просуммируем полученные равенства:

$$\sum_{k=0}^{T-1} \langle -\nabla V_{x_k}(x_{k+1}), x_{k+1} - y\rangle = V_{x_0}(y) - V_{x_T}(y) - \sum_{k=0}^{T-1} V_{x_k}(x_{k+1})$$

Имея ввиду полученное уравнение, давайте, наконец, попробуем сформулировать метод **зеркального спуска**:

$$x_{k+1} = \arg\min_{x\in S} \left(\langle \alpha_k g_k, x\rangle + V_{x_k}(x)\right)$$

Посмотрим внимательнее на условие проекции для точки x_{k+1}:

$$\langle \alpha_k g_k, x_{k+1} - y\rangle + \langle \nabla V_{x_k}(x_{k+1}), x_{k+1} - y\rangle \leqslant 0$$

$$\langle \alpha_k g_k, x_{k+1} - y \rangle \leqslant -\langle \nabla V_{x_k}(x_{k+1}), x_{k+1} - y \rangle$$

Попробуем теперь получить наше базовое неравенство, используя (baseMD):

$$\begin{aligned}
\langle \alpha_k g_k, x_k - y \rangle &\leqslant -\langle \nabla V_{x_k}(x_{k+1}), x_{k+1} - y \rangle - \langle \alpha_k g_k, x_{k+1} - x_k \rangle = \\
&= V_{x_k}(y) - V_{x_{k+1}}(y) - V_{x_k}(x_{k+1}) - \langle \alpha_k g_k, x_{k+1} - x_k \rangle \leqslant \\
&\leqslant V_{x_k}(y) - V_{x_{k+1}}(y) - \frac{1}{2}\|x_k - x_{k+1}\|^2 - \langle \alpha_k g_k, x_{k+1} - x_k \rangle \leqslant \\
&\leqslant V_{x_k}(y) - V_{x_{k+1}}(y) + \left(\langle \alpha_k g_k, x_k - x_{k+1} \rangle - \frac{1}{2}\|x_k - x_{k+1}\|^2 \right) \leqslant \\
&\leqslant V_{x_k}(y) - V_{x_{k+1}}(y) + \frac{\alpha_k^2}{2}\|g_k\|_*^2
\end{aligned}$$

ТЕЛЕСКОПИРУЕМ

$$\begin{aligned}
\sum_{k=0}^{T-1} \langle \alpha_k g_k, x_k - y \rangle &\leqslant V_{x_0}(y) - V_{x_T}(y) + \sum_{k=0}^{T-1} \frac{\alpha_k^2}{2}\|g_k\|_*^2 \\
&\leqslant V_{x_0}(y) + \sum_{k=0}^{T-1} \frac{\alpha_k^2}{2}\|g_k\|_*^2 \\
&\leqslant M + \frac{\alpha^2 G^2 T}{2}
\end{aligned}$$

Здесь мы подразумеваем $\|g_k\|_* \leqslant G$ равномерно по k, а $V_{x_0}(y) \leqslant M$. В итоге:

$$\begin{aligned}
f(\overline{x}) - f^* = f\left(\frac{1}{T}\sum_{k=0}^{T-1} x_k \right) - f^* &\leqslant \frac{1}{T}\left(\sum_{k=0}^{T-1} f(x_k) - f^* \right) \\
&\leqslant \frac{1}{T}\left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \\
&\leqslant \frac{M}{\alpha T} + \frac{\alpha G^2}{2} \leqslant \sqrt{\frac{2MG^2}{T}}
\end{aligned}$$

Выбирая шаг $\alpha_k = \alpha = \sqrt{\dfrac{2M}{G^2 T}}$

2.4 Алгоритм зеркального спуска (mirror descent):

$$x_{k+1} = \arg\min_{x \in S} \left(\langle \alpha_k g_k, x \rangle + V_{x_k}(x) \right)$$

Интересные фишки:

- Такая же скорость сходимости, как и для метода проекции субградиента.
- Работает в существенно более широком классе практических задач.

2.5 Ценность

В случае, если мы выбрали $\| \cdot \| = \| \cdot \|_2$ Евклидову норму и Евклидово расстояние, то этот метод в точности совпадает с тем, что мы уже называем метод проекции субградиента.

Значит, надо предоставить сценарий, когда МЗС работает лучше, давайте рассмотрим $S = \Delta_n$ - вероятностный симплекс, а так же следующее расстояние Брэгмана $V_x(y) = \sum_{i \in [n]} y_i \log \frac{y_i}{x_i} = D(y \| x)$. Норма в прямом пространстве при этом $\|\|_1$, а в сопряженном - $\|\|_\infty$. Кроме того, для заданной дивергенции Брэгмана справедливо:

$$x_0 = (1/n, \dots, 1/n) \ \rightarrow \ V_{x_0}(x) \leqslant \log n \ \ \forall x \in \Delta_n$$

Тогда алгоритм зеркального спуска запишется в виде:

$$\begin{aligned} x_{k+1} &= \arg\min_{x \in S} \left(\langle \alpha_k g_k, x \rangle + V_{x_k}(x) \right) \\ &= \arg\min_{x \in S} \left(\langle \alpha_k g_k, x \rangle + \sum_{i \in [n]} x_i \log \frac{x_i}{x_{k_i}} \right) \\ &= x_k \cdot \frac{e^{-\alpha_k g_k}}{\| x_k \cdot e^{-\alpha_k g_k} \|_1} \end{aligned}$$

А оценки с учетом того, что $M = \log n$, $\|g_k\|_\infty \leqslant G$ запишутся, как:

$$f(\overline{x}) - f^* \leqslant \sqrt{\frac{2 \log(n) G^2}{T}}$$

Таким образом, получаем алгоритм зеркального спуска:

Algorithm 1 Mirror Descent

Input: $x_0, f(x), \nabla f(x), \alpha$

for $k = 0{,}1, \ldots ,K$ **do**

  $x_{k+1} = \arg\min\limits_{x \in S} \left(\langle \alpha_k g_k, x\rangle + V_{x_k}(x)\right)$

end for

Output: x_{K+1}

3 Описание методов

3.1 Двойственные усреднения

Предлагается исследовать также метод двойственных усреднений. В общем виде:

Algorithm 2 Dual Averaging

Input: Set $s_0 = 0$,Choose $\beta_0 > 0$.

for $k = 0{,}1, \ldots ,K$ **do**

  Compute $g_k = \nabla f(x_k)$

  Choose $\lambda_k > 0$. Set $s_{k+1} = s_k + \lambda_k g_k$

  Choose $\beta_{k+1} \geqslant \beta_k$. Set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$.

end for

Где $\pi_\beta(s) = \arg\min\limits_{x \in Q}(-\langle s, x\rangle + \beta d(x))$, причем $d(x) = \sum\limits_i x_i \log x_i$.

Тогда, получаем:

$$
\begin{aligned}
&\arg\min_{x \in Q}(-\langle s, x\rangle + \beta \sum_i x_i \log x_i) \\
&= \arg\min(\langle \frac{-s}{\beta}, x\rangle + \sum_i x_i \log \frac{x_i}{x_0} + \sum_i x_i \log x_0) \\
&= \arg\min(\langle \frac{-s}{\beta}, x\rangle + \sum_i x_i \log \frac{x_i}{x_0} - \sum_i x_i \log n) \\
&= \arg\min(\langle \frac{-s}{\beta}, x\rangle + \sum_i x_i \log \frac{x_i}{x_0} - \log n) \\
&= \arg\min(\langle \frac{-s}{\beta}, x\rangle + \sum_i x_i \log \frac{x_i}{x_0} \\
&= x_0 \cdot \frac{e^{\frac{s}{b}}}{\|x_0 \cdot e^{\frac{s}{b}}\|} \\
&= \frac{e^{\frac{s}{b}}}{\|e^{\frac{s}{b}}\|}
\end{aligned}
$$

А значит, можно выразить $x_{k+1} = x_k \cdot \frac{e^{-g_k \alpha}}{\|x_k e^{-g_k \alpha}\|}$.

3.2 Метод подобных треугольников

Заметим, что оба, описанных выше метода могут быть ускорены с помощью метода подобных треугольников.

Algorithm 3 Similar Triangles

$y_{k+1} = \frac{\alpha_{k+1} u_k + A_k x_k}{A_{k+1}}$
$u_{k+1} = \arg\min\limits_{x \in Q} \varphi_{k+1}(x)$
$x_{k+1} = \frac{\alpha_{k+1} u_{k+1} + A_k x_k}{A_{k+1}}$
$A_k = \sum\limits_{i=0}^{k} \alpha_i, \alpha_0 = L^{-1}, x_0 = u_0 = \arg\min\limits_{x \in Q} \varphi_0(x)$

Где α_{k+1} можно задавать рекуррентно:

$$\alpha_{k+1} = \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \alpha_k^2},$$

и тогда мы получим еще 2 метода TR-MD и TR-DA.

Но можно подойти к выбору α_k чуть более нестандартно.

Возьмем:

$$\alpha_k = \arg\min_{\alpha_k \in M}(f(y_k)) = \arg\min_{\alpha_k \in M} f(\alpha_{k+1} u_k + (1 - \alpha_{k+1}) x_k)$$

Таким образом вывели еще два оптимизированных методом подобных треугольников, но для других альфа алгоритма TR-MD2 и TR-DA2.

4 Эксперименты

В ходе экспериментов мы тестируем работу полученных выше алгоритмов и сравниваем скорость их сходимости к точному решению. Для тестирования взяли функцию:

$$f(x) = \|Ax - b\|_2^2$$

$$\nabla f(x) = 2A^T(Ax - b)$$

И будем искать ее минимум на выпуклом множестве $S : \left\{ x \in S : x \geqslant 0, \sum\limits_{i \in [n]} x_i = 1 \right\}$ (вероятностном симплексе).

Но для последних 2-х методов знания значений градиента оказывается недостаточно. Необходимо в явном виде выразить α.

$$\alpha = \arg\min_{\alpha} f(\alpha x^n + (1-\alpha) y^n)$$

$$\sum_{k=1}^{n} (x_k^n - y_k^n) \frac{\partial f(\alpha x^n + (1-\alpha) y^n)}{\partial x_k}$$

$$(x^n - y^n)^T \nabla f(\alpha x^n + (1-\alpha) y^n) = 0$$

$$(x^n - y^n)^T 2A^T (A(\alpha x^n + (1-\alpha) y^n) - b) = 0$$

$$2(x^n - y^n)^T (\alpha (A^T (A x^n - b)) + (1-\alpha)(A^T (A y^n - b))) = 0$$

$$(x^n - y^n)^T (\alpha \nabla f(x^n) + (1-\alpha) \nabla f(y^n)) = 0$$

$$\alpha = \begin{cases} const = 0 & (x^n - y^n)^T (\nabla f(x^n) - \nabla f(y^n)) = 0 \\ -\frac{(x^n - y^n)^T (\nabla f(y^n))}{(x^n - y^n)^T (\nabla f(x^n) - \nabla f(y^n))} & else \end{cases}$$

Полученные результаты представлены на графиках.

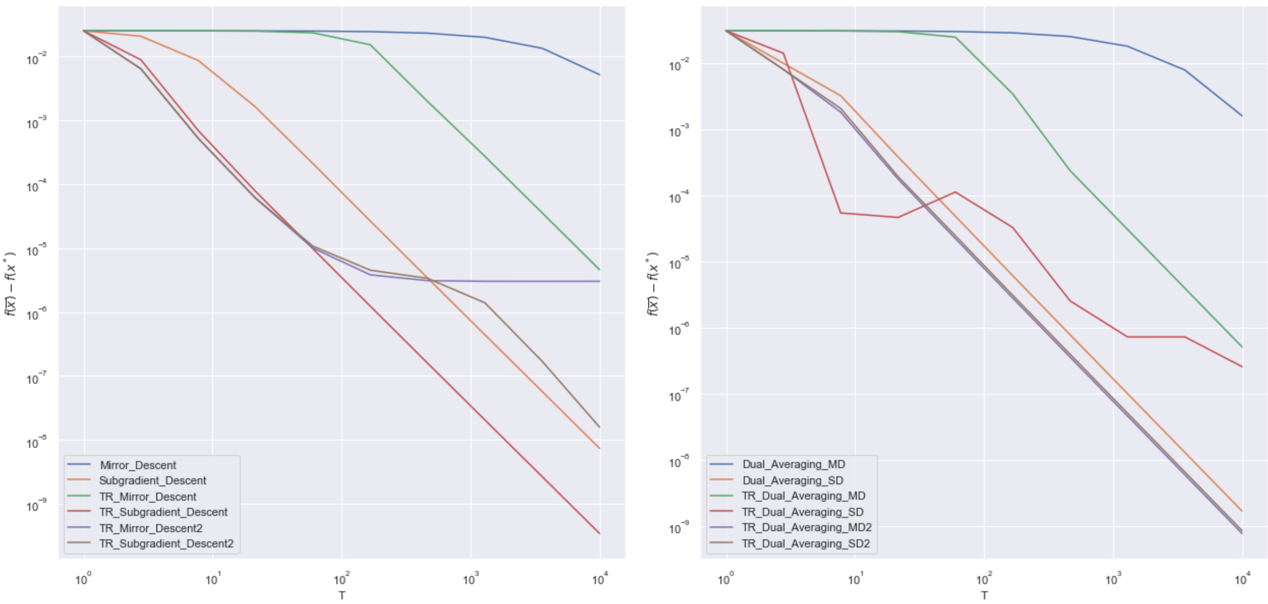

Рис. 1: На левом графике результаты работы для метода зеркального спуска и его ускорений, на правом аналогично для метода двойственных усреднений.

5 Заключение

В данной работе был реализован метод зеркального спуска и рассмотрены различные его вариации. Нетрудно заметить из графиков, что ускорение получается более чем существенное и ускоренные методы сходятся значительно быстрее. Наиболее быстрым оказывается метод двойственных с взвешенными λ оптимизированный методом подобных треугольников с особым выбором коэффициента α.

Responsibilities:

- Code director - Якушев А.Ю.
- Text director - Ефремов С.В.

Список литературы

[1] *Гасников А.В.* Современные численные методы оптимизации. Метод универсального градиентного спуска

[2] *Камзолов Д.И.* Семинары 674 группы

[3] *Камзолов Д.И.* Презентации 674 группы

[4] *Нестеров Ю.* Primal-dual subgradient methods for convex problems

[5] *Гасников А.В., Нестеров Ю.Е.* Универсальный метод для задач стохастической композитной оптимизации

[6] *Меркулов Д.* Презентации