Метод зеркального спуска

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

21 ноября 2023

ФПМИ

Историческая мотивация

- Посмотрим на итерацию градиентного спуска:

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

- Пусть x принадлежит банахову пространству $(E, \|\cdot\|)$. **Вопрос:** а что можем сказать про $\nabla f(x^k)$? В общем случае $\nabla f(x^k)$ лежит не в $(E, \|\cdot\|)$, а в $(E^*, \|\cdot\|_*)$. Тогда с чего мы вдруг начали складывать векторы из абсолютно разных пространств...
- Ничего страшного тут нет в случае когда $\|\cdot\| = \|\cdot\|_2$, тогда $(E^*, \|\cdot\|_*) = (E, \|\cdot\|)$ и все что мы делали было валидно.
- Хотим попробовать выйти за пределы евклидовости. Расстояние не обязательно мерять в евклидовой норме (не смотря на то, что в конечномерном случае все нормы эквивалентны). «Геометрия» задачи может подталкивать использовать другие способы измерения расстояния. Зачем, например, мерять расстояние между распределениями вероятности в евклидовой норме, есть более «физичные» способы.

Историческая мотивация

- А. Немировский и Д. Юдин:

$$\varphi(x^{k+1}) = \varphi(x^k) - \gamma \nabla f(x^k)$$

  где φ подбирается так, что φ действует из E в E^* и φ^{-1} из E^* в E.
- Получается мы переходим в «зеркальное» пространство E^*, там делаем шаг градиентного спуска, а потом с помощью φ^{-1} возвращаемся к x^{k+1} из E.
- Это и есть идея «зеркальности» градиентного спуска.

Дивергенция Брэгмана

Определение μ-сильной выпуклости

Пусть дана непрерывно дифференцируемая на выпуклом множестве $\mathcal{X}$ функция $d : \mathcal{X} \to \mathbb{R}$. Будем говорить, что она является μ-сильно выпуклой ($\mu > 0$) относительно нормы $\|\cdot\|$ на множестве $\mathcal{X}$, если для любых $x, y \in \mathcal{X}$ выполнено

$$d(x) \geq d(y) + \langle \nabla d(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2.$$

Напоминаем, что в курсе мы определяем выпуклость функции только на выпуклых множествах $\mathcal{X}$.

Дивергенция Брэгмана

Определение

Пусть дана дифференцируемая 1-сильно выпуклая относительно нормы $\|\cdot\|$ на множестве $\mathcal{X}$ функция d. Дивергенцией Брэгмана, порожденной функцией d на множестве $\mathcal{X}$, называется функция двух аргументов $V(x,y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ такая, что для любых $x, y \in \mathcal{X}$

$$V(x,y) = d(x) - d(y) - \langle \nabla d(y), x - y \rangle.$$

Дивергенцию Брэгмана можно определять и для строго выпуклых функций.

Дивергенция Брэгмана: примеры

- $d(x) = \frac{1}{2}\|x\|_2^2$ на $\mathbb{R}^d$. **Вопрос:** какую дивергенцию породит эта функция d? $V(x, y) = \frac{1}{2}\|x - y\|_2^2$.
- $d(x) = \sum_{i=1}^d x_i \log x_i$ на вероятностном симплексе $\triangle_d = \{x \in \mathbb{R}^d \mid x_i \geq 0, \sum_{i=1}^d x_i = 1\}$. Неравенство Пинскера гарантирует 1-сильную выпуклость относительно $\|\cdot\|_1$. $V(x, y) = \sum_{i=1}^d x_i \log \frac{x_i}{y_i}$ (KL-дивергенция).
- $d(X) = \mathrm{trace}(X \log X)$. Квантовая дивергенция фон Неймана $V(X, Y) = \mathrm{trace}(X \log X - X \log Y - X + Y)$
- $d(X) = -\log(\det X)$. $V(X, Y) = \mathrm{trace}(XY^{-1} - I) - \log\det(XY^{-1})$

Дивергенция Брэгмана: свойства

- Ассиметричность - смотри KL-дивергенцию
- Сильная выпуклость дает важное свойство (напрямую из определения)

Свойство дивергенции Брэгмана

Для любых точек $x, y \in \mathcal{X}$ следует что $V(x, y) \geq \frac{1}{2}\|x - y\|^2$.

- Отсюда вытекает сразу неотрицательность.
- Невыпукла по второму аргументу.

Дивергенция Брэгмана: свойства

Равенство параллелограмма/теорема Пифагора

Для любых точек $x, y, z \in \mathcal{X}$ следует что

$$V(z,x) + V(x,y) - V(z,y) = \langle \nabla d(y) - \nabla d(x), z - x \rangle.$$

Доказательство

По определению:

$$
\begin{aligned}
V(z,x) + V(x,y) =& d(z) - d(x) - \langle \nabla d(x), z - x \rangle \\
& + d(x) - d(y) - \langle \nabla d(y), x - y \rangle \\
=& d(z) - d(y) - \langle \nabla d(y), z - y \rangle \\
& - \langle \nabla d(x) - \nabla d(y), z - x \rangle \\
=& V(z,y) - \langle \nabla d(x) - \nabla d(y), z - x \rangle .
\end{aligned}
$$

А это то, что нужно.

Метод зеркального спуска

Решаем задачу

$$\min_{x \in \mathcal{X}} f(x),$$

где множество $\mathcal{X}$ и функция f выпуклы.

- Метод зеркального спуска:

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \{\langle \gamma \nabla f(x^k), x \rangle + V(x, x^k)\}$$

- **Вопрос:** если $d(x) = \frac{1}{2}\|x\|_2^2$, то какой метод получится? Градиентный спуск с евклидовой проекцией.
- **Вопрос:** если возьмем $\mathcal{X} = \mathbb{R}^d$ и некоторую d, как будет выглядеть условие оптимальности для шага метода?

$$\gamma \nabla f(x^k) + \nabla d(x^{k+1}) - \nabla d(x^k) = 0$$

Метод зеркального спуска

- С предыдущего слайда:

$$\nabla d(x^{k+1}) = \nabla d(x^k) - \gamma \nabla f(x^k)$$

Это и есть идея Немировского и Юдина! Идея "зеркальности".

- ∇d переносит нас из E в E^*, там мы можем оперировать с $\nabla f(x^k)$. Сделаем шаг градиентного спуска в "зеркальном" пространстве и получим некоторый вектор $\nabla d(x^k) - \gamma \nabla f(x^k)$. С помощью $(\nabla d)^{-1}$ можно получить x^{k+1}:

$$x^{k+1} = (\nabla d)^{-1}(\nabla d(x^k) - \gamma \nabla f(x^k))$$

- В жизни будет все проще. У $\arg\min$ либо есть явное аналитическое решение, либо его можно отрешать методом оптимизации до хорошей точности.

Гладкость: определение

Определение L-гладкой функции

Пусть дана непрерывно дифференцируемая на $\mathcal{X}$ функция $f : \mathcal{X} \to \mathbb{R}$. Будем говорить, что данная функция имеет L-Липшицев градиент (говорить, что она является L-гладкой) относительно нормы $\|\cdot\|$ на $\mathcal{X}$, если для любых $x, y \in \mathcal{X}$ выполнено

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|.$$

Обобщение гладкости на произвольную норму.

Гладкость: свойства

Теорема (свойство L - гладкой функции)

Пусть дана L - гладкая относительно нормы $\|\cdot\|$ функция $f : \mathcal{X} \to \mathbb{R}$. Тогда для любых $x, y \in \mathcal{X}$ выполнено

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|x - y\|^2.$$

Гладкость: свойства

Доказательство

Начнем с формулы Ньютона-Лейбница

$$f(y) - f(x) = \int\limits_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle d\tau$$

$$= \langle \nabla f(x), y - x \rangle + \int\limits_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau$$

Тогда

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| = \left| \int\limits_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau \right|$$

$$\leq \int\limits_0^1 |\langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle| d\tau$$

Гладкость: свойства

Доказательство

Применим КБШ ($\langle x, y\rangle \leq \|x\| \cdot \|y\|_*$):

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \int\limits_0^1 |\langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x\rangle| d\tau$$

$$\leq \int\limits_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_* \|y - x\| d\tau$$

Далее определение L-гладкости относительно нормы $\|\cdot\|$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq L\|y - x\|^2 \int\limits_0^1 \tau d\tau$$

$$= \frac{L}{2}\|x - y\|^2$$

Доказательство сходимости зеркального спуска

- Условия оптимальности для шага зеркального спуска для любого $x \in \mathcal{X}$:

$$\langle \gamma \nabla f(x^k) + \nabla d(x^{k+1}) - \nabla d(x^k), x^{k+1} - x \rangle \leq 0$$

- Свойство дивергенции Брэгмана $(V(z, x) + V(x, y) - V(z, y) = \langle \nabla d(x) - \nabla d(y), x - z \rangle)$:

$$\gamma \langle \nabla f(x^k), x^{k+1} - x \rangle + V(x, x^{k+1}) + V(x^{k+1}, x^k) - V(x, x^k) \leq 0$$

Доказательство сходимости зеркального спуска

- С прошлого слайда:

$$\gamma\langle \nabla f(x^k), x^{k+1} - x\rangle + V(x, x^{k+1}) + V(x^{k+1}, x^k) - V(x, x^k) \leq 0$$

- Гладкость:

$$f(x^{k+1}) - f(x^k) - \langle \nabla f(x^k), x^{k+1} - x^k\rangle \leq \frac{L}{2}\|x^k - x^{k+1}\|^2$$

- Складываем, домножив второе на $\gamma > 0$:

$$\gamma\langle \nabla f(x^k), x^k - x\rangle + \gamma f(x^{k+1}) - \gamma f(x^k)$$
$$+ V(x, x^{k+1}) + V(x^{k+1}, x^k) - V(x, x^k) \leq \frac{\gamma L}{2}\|x^k - x^{k+1}\|^2$$

Доказательство сходимости зеркального спуска

- Немного перегруппируем:

$$
\begin{aligned}
&\gamma\langle \nabla f(x^k), x^k - x\rangle + \gamma\left(f(x^{k+1}) - f(x^k)\right) \\
&\leq V(x, x^k) - V(x, x^{k+1}) + \frac{\gamma L}{2}\|x^k - x^{k+1}\|^2 - V(x^{k+1}, x^k)
\end{aligned}
$$

- Выпуклость f:

Доказательство сходимости зеркального спуска

- Немного перегруппируем:

$$\gamma\langle \nabla f(x^k), x^k - x\rangle + \gamma\left(f(x^{k+1}) - f(x^k)\right)$$
$$\leq V(x, x^k) - V(x, x^{k+1}) + \frac{\gamma L}{2}\|x^k - x^{k+1}\|^2 - V(x^{k+1}, x^k)$$

- Выпуклость f:

$$\gamma\left(f(x^{k+1}) - f(x)\right)$$
$$\leq V(x, x^k) - V(x, x^{k+1}) + \frac{\gamma L}{2}\|x^k - x^{k+1}\|^2 - V(x^{k+1}, x^k)$$

Доказательство сходимости зеркального спуска

- Немного перегруппируем:

$$\gamma\langle\nabla f(x^k), x^k - x\rangle + \gamma\left(f(x^{k+1}) - f(x^k)\right)$$
$$\leq V(x, x^k) - V(x, x^{k+1}) + \frac{\gamma L}{2}\|x^k - x^{k+1}\|^2 - V(x^{k+1}, x^k)$$

- Выпуклость f:

$$\gamma\left(f(x^{k+1}) - f(x)\right)$$
$$\leq V(x, x^k) - V(x, x^{k+1}) + \frac{\gamma L}{2}\|x^k - x^{k+1}\|^2 - V(x^{k+1}, x^k)$$

- Свойство дивергенции $\frac{1}{2}\|x^k - x^{k+1}\|^2 \leq V(x^{k+1}, x^k)$:

$$\gamma\left(f(x^{k+1}) - f(x)\right) \leq V(x, x^k) - V(x, x^{k+1}) + (\gamma L - 1)\, V(x^{k+1}, x^k)$$

Доказательство сходимости зеркального спуска

- $\gamma \leq 1/L$:

$$\gamma \left(f(x^{k+1}) - f(x) \right) \leq V(x, x^k) - V(x, x^{k+1})$$

- Суммируя по всем k от 0 до $K-1$:

$$\frac{\gamma}{K} \sum_{k=0}^{K-1} \left(f(x^{k+1}) - f(x) \right) \leq \frac{1}{K} \sum_{k=0}^{K-1} \left(V(x, x^k) - V(x, x^{k+1}) \right)$$

- Получаем:

$$\frac{1}{K} \sum_{k=1}^{K} \left(f(x^k) - f(x) \right) \leq \frac{1}{\gamma K} \left(V(x, x^0) - V(x, x^K) \right) \leq \frac{V(x, x^0)}{\gamma K}$$

Доказательство сходимости зеркального спуска

- Неравенство Йенсена:

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f(x) \leq \frac{V(x, x^0)}{\gamma K}$$

- Подставляем $x = x^*$:

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f(x^*) \leq \frac{V(x^*, x^0)}{\gamma K}$$

Сходимость зеркального спуска

Теорема сходимость зеркального спуска для L-гладких относительно $\|\cdot\|$ и выпуклых функций

Пусть задача оптимизации на выпуклом множестве $\mathcal{X}$ с L-гладкой относительно нормы $\|\cdot\|$, выпуклой целевой функцией f решается с помощью зеркального спуска с шагом $\gamma \leq \frac{1}{L}$. Тогда справедлива следующая оценка сходимости

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f(x^*) \leq \frac{V(x^*, x^0)}{\gamma K}$$

Сходимость зеркального спуска

- Результат очень похож на сходимость градиентного спуска и обобщает его.
- **Вопрос:** а может ли оценка зеркального спуска быть лучше, чем для градиентного? Где проявится этот эффект? В L и V.
- Гладкость, которую использовали сегодня
$$\|\nabla f(x) - \nabla f(y)\|_q \leq L\|x - y\|_p, \quad \frac{1}{p} + \frac{1}{q} = 1.$$
- Если $1 \leq p \leq 2$, то $\| \cdot \|_2 \leq \| \cdot \|_p, \quad \| \cdot \|_q \leq \| \cdot \|_2$.
- **Вопрос:** что тогда можно сказать про отношение L и L_2(которую использовали ранее в евклидовом случае)? $L \leq L_2$, а в хорошем случае L значительно меньше.
- **Вопрос:** с дивергенций тоже все хорошо? Там ситуация обратная для $1 \leq p \leq 2$, $V(x,y) \geq \frac{1}{2}\|x - y\|_p^2 \geq \frac{1}{2}\|x - y\|_2^2$
- Выигрыш будет, если $\frac{L_2}{L}$ значительно больше, чем $\sup_{x,y\in\mathcal{X}} \frac{2V(x,y)}{\|x-y\|_2^2}$. Следующий пример из таких.

Пример зеркального спуска

$$x^{k+1} = \arg\min_{x\in\mathcal{X}}\{\langle\gamma\nabla f(x^k), x\rangle + V(x, x^k)\}$$

Найдем явный вид для $V(x,y) = \sum_{i=1}^d x_i \log\left(\frac{x_i}{y_i}\right)$ на $\triangle_d$.

- Формальная запись задачи минимизации:

$$\min_{x\in\triangle_d} \langle\gamma\nabla f(x^k), x\rangle + V(x, x^k)$$

$$\text{s.t.} \quad -x_i \leq 0$$

$$\sum_{i=1}^d x_i - 1 = 0$$

- Лагранжиан:

$$L(x,\lambda,\nu) = \langle\gamma\nabla f(x^k), x\rangle + V(x, x^k) + \sum_{i=1}^d \lambda_i(-x_i) + \nu\left(\sum_{i=1}^d x_i - 1\right)$$

Пример зеркального спуска

- Распишем покомпонентно:

$$L(x,\lambda,\nu) = \langle \gamma \nabla f(x^k), x\rangle + V(x,x^k) + \sum_{i=1}^d \lambda_i(-x_i) + \nu\left(\sum_{i=1}^d x_i - 1\right)$$

$$= \sum_{i=1}^d (\log\left(\frac{x_i}{x_i^k}\right) + \gamma[\nabla f(x^k)]_i - \lambda_i + \nu)x_i - \nu$$

- Минимизируем по каждой x_i, чтобы получить двойственную:

$$\inf_x L(x,\lambda,\nu) = \sum_{i=1}^d -x_i^k \exp(-1 + \lambda_i - \gamma[\nabla f(x^k)]_i - \nu) - \nu$$

- Двойственная задача:

$$\max_{\lambda_i \geq 0, \nu \in \mathbb{R}} \left[\sum_{i=1}^d -x_i^k \exp(-1 + \lambda_i - \gamma[\nabla f(x^k)]_i - \nu) - \nu\right]$$

Пример зеркального спуска

- Двойственная задача:

$$\max_{\lambda_i \geq 0, \nu \in \mathbb{R}} \left[\sum_{i=1}^{d} -x_i^k \exp(-1 + \lambda_i - \gamma[\nabla f(x^k)]_i - \nu) - \nu \right]$$

- **Вопрос:** что можем сказать про λ_i^*? Видно, что $\lambda_i^* = 0$. Это все что нужно было от двойственной.
- Условие ККТ (здесь сразу подставлены $\lambda_i^* = 0$):

$$\nabla_x \left(\langle \gamma \nabla f(x^k), x \rangle + V(x, x^k) + \nu \left(\sum_{i=1}^{d} x_i - 1 \right) \right) = 0$$

- Откуда

$$\log\left(\frac{x_i^*}{x_i^k}\right) + 1 + \gamma[\nabla f(x^k)]_i + \nu^* = 0$$

Пример зеркального спуска

- Преобразуем и получаем:

$$x_i^* = x_i^k \exp(-\gamma[\nabla f(x^k)]_i) \cdot \exp(1 + \nu^*)$$

- **Вопрос:** из каких соображений подобрать ν^*? $\sum_{i=1}^d x_i^* = 1$, тогда окончательно:

$$x_i^{k+1} = x_i^* = \frac{x_i^k \exp(-\gamma[\nabla f(x^k)]_i)}{\sum_{i=1}^d x_i^k \exp(-\gamma[\nabla f(x^k)]_i)}.$$

  Это и есть итерации зеркального спуска для симплекса.
- В случае симплекса и KL-дивергенции можно получить выигрыш в $\frac{d}{\log d}$ раз по сравнению с градиентным спуском с евклидовой проекцией.