Метод зеркального спуска

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

26 октября 2023

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 $\nabla f(x^k)$?

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...

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.

$\| \, \|_p \to \| \, \|_* = \| \, \|_q$ $\quad \frac{1}{p} + \frac{1}{q} = 1$

$\| \, \|_2 \quad \| \, \|_2 = \| \, \|_*$

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:

$$\varphi(x^{k+1}) = \underbrace{\varphi(x^k)}_{\in E_*} - \underline{\gamma \nabla f(x^k)}$$

где подбирается так, что φ действует из E в $E^{\cancel{*}}_{*}$ и φ^{-1} из $E^{\cancel{*}}_{*}$ в E.

Историческая мотивация

- А. Немировский и Д. Юдин:

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

где 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} вернутся к x^{k+1} из E.

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

Историческая мотивация

- А. Немировский и Д. Юдин:

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

  где 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} вернутся к x^{k+1} из E.
- Это и есть идея «зеркальности» градиентного спуска.

Дивергенция Брэгмана

Определение μ-сильной выпуклости

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $d : \mathbb{R}^d \to \mathbb{R}$. 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

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

Дивергенция Брэгмана

Определение μ-сильной выпуклости

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $d : \mathbb{R}^d \to \mathbb{R}$. 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.

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 $x, y \in \mathcal{X}$

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

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.

$$d(x) = \frac{1}{2}\|x\|_2^2 \qquad V(x,y) =$$

$$= \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y; x-y \rangle$$

$$= \frac{1}{2}\left(\|x\|_2^2 - \|y\|_2^2 - 2\langle y; x\rangle + 2\|y\|_2^2\right) = \frac{1}{2}\left(\|x\|_2^2 + \|y\|_2^2 - 2\langle x, y\rangle\right)$$

$$= \frac{1}{2}\|x-y\|_2^2$$

Дивергенция Брэгмана: примеры

- $d(x) = \frac{1}{2}\|x\|_2^2$ на $\mathbb{R}^d$. **Вопрос:** какую дивергенцию породит эта функция d?

Дивергенция Брэгмана: примеры

- $d(x) = \frac{1}{2}\|x\|_2^2$ на $\mathbb{R}^d$. **Вопрос:** какую дивергенцию породить эта функция d? $V(x, y) = \frac{1}{2}\|x - y\|_2^2$.

Дивергенция Брэгмана: примеры

- $d(x) = \frac{1}{2}\|x\|_2^2$ на $\mathbb{R}^d$. **Вопрос:** какую дивергенцию породить 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 $\triangle_d = \{x \in \mathbb{R}^d \mid x_i \geq 0, \sum_{i=1}^{d} x_i = 1\}$.

Дивергенция Брэгмана: примеры

- $d(x) = \frac{1}{2}\|x\|_2^2$ на $\mathbb{R}^d$. **Вопрос:** какую дивергенцию породить 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).

Дивергенция Брэгмана: примеры

- $d(x) = \frac{1}{2}\|x\|_2^2$ на $\mathbb{R}^d$. **Вопрос:** какую дивергенцию породить 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 $\boxed{V(x,y) \geq \frac{1}{2}\|x-y\|^2.}$

- 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:

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

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:

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

Метод зеркального спуска

Решаем задачу

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

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

$x^k - \gamma\nabla f(x^k) = y^k$

$\arg\min_{x\in X} \| y^k - x\|_2^2$

$= \arg\min_{x\in X} \{ \| x^k - \gamma \nabla f(x^k) - x\|_2^2 \}$

- Метод зеркального спуска:

$$x^{k+1} = \arg\min_{x\in\mathcal{X}}\{2\langle\gamma\nabla f(x^k), x\rangle + V(x, x^k)\}$$

$+ \quad \|x - x^k\|_2^2$

- **Вопрос:** если $d(x) = \frac{1}{2}\|x\|_2^2$, то какой метод получится?

$2\langle \gamma \nabla f(x^k); x\rangle + \|x - x^k\|_2^2 + \gamma^2 \|\nabla f(x^k)\|_2^2$

$-2\langle \gamma \nabla f(x^k); x^k\rangle$

$= \quad \|x^k - \gamma \nabla f(x^k) - x\|_2^2$

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.

Метод зеркального спуска

Решаем задачу

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

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

- Метод зеркального спуска:

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

$\mathbb{R}^d$

$d(x) - d(x^k) - \langle \nabla d(x^k); x - x^k \rangle$

$d(x) - \langle \nabla d(x^k); x \rangle$

- **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, когда будет выглядеть условие оптимальности для шага метода?

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

Метод зеркального спуска

Решаем задачу

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

$\langle \nabla g(x^{k+1}); x^{k+1} - x \rangle \geq 0$

$\forall x \in X$

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

- Метод зеркального спуска:

$g(x)$

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

$\langle \nabla g(x^*); x^* - x \rangle \geq 0$

- **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, когда 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:

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

  Это и есть идея Немировского и Юдина! Идея "зеркальности".
- ∇d переносит нас из E в E^*, там мы можем оперировать с $\nabla f(x^k)$.

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 $\nabla d(x^k) - \gamma \nabla f(x^k)$.

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}:

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

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

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

2 2

Гладкость: определение

Определение L-гладкой функции

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

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

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

Гладкость: свойства

Доказательство

Начнем с формулы Ньютона-Лейбница

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle d\tau$$

$$= \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau$$

Тогда

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| = \left| \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau \right|$$

$$\leq \int_0^1 |\langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle| d\tau$$

Гладкость: свойства

Доказательство

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

$\langle x; y \rangle \leq \|x\|_2 \|y\|_2$

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

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

Гладкость: свойства

Доказательство

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

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

$C_1\| \ \|_t \leq \| \ \|_p \leq C_2\| \ \|_t$

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

$\|x\|_p \geq \|x\|_q$

$p < q$

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

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

$\|x\|_1 \geq \|x\|_2 \geq \|x\|_\infty$

$\max|x_i|$

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

$\sqrt{d} \geq 1$

Доказательство сходимости зеркального спуска

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

Доказательство сходимости зеркального спуска

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

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

$$g(x) = \langle \gamma \nabla f(x^k); x \rangle + d(x) - \langle \nabla d(x^k); x \rangle$$

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

$$\langle \nabla d(x^{k+1}) - \nabla d(x^k); x^{k+1} - x \rangle =$$

$$= V(x^{k+1}; x) + V(x; x^k) - V(x^{k+1}; x^k)$$

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

x^{k+1} x^k x^{k+1} x

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

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

$V(x; x^{k+1}) + V(x^{k+1}; x^k) - V(x; x^k)$

Доказательство сходимости зеркального спуска

- С прошлого слайда:

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

$\langle \gamma \nabla f(x^k); x^k - x \rangle$

- 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 $\gamma > 0$:

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

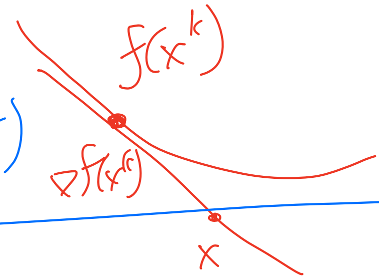
$$\gamma\langle \nabla f(x^k); x^k - x\rangle + \gamma\left(f(x^{k+1}) - f(x^k)\right)$$

$$\leq \gamma\frac{L}{2}\|x^k - x^{k+1}\|^2 + V(x; x^k) - V(x; x^{k+1}) - V(x^{k+1}; x^k)$$

$$\langle \nabla f(x^k); x^k - x\rangle - f(x^k) \geq -f(x)$$

$$f(x^k) + \langle \nabla f(x^k); x - x^k\rangle \leq f(x)$$

$$-f(x^k) - \langle \nabla f(x^k); x^k - x\rangle \geq -f(x)$$



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

$$\leq \gamma\frac{L}{2}\|x^k - x^{k+1}\|^2 + V(x; x^k) - V(x; x^{k+1}) - V(x^{k+1}; x^k)$$

$$-V(x^{k+1}; x^k) \leq -\frac{1}{2}\|x^{k+1} - x^k\|^2$$

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

$\|x^k - x\|_2^2$ $\qquad$ $\|x^{k+1} - x\|_2^2$

$$\le \quad V(x; x^k) - V(x; x^{k+1})$$

$$+ \underbrace{\left(-\frac{1}{2} + \frac{\gamma L}{2}\right)\|x^k - x^{k+1}\|^2}_{\le 0}$$

$$\gamma \le \frac{1}{L}$$

$$\gamma\left(f(x^{k+1}) - f(x)\right) \le \underline{V(x; x^k) - V(x; x^{k+1})}$$

$$\frac{1}{k}\sum$$

$$\underbrace{\frac{1}{k}\sum f(x^{k+1})}\ - f(x) \le \frac{1}{\gamma k}\left(V(x, x^0) - \underbrace{\cancel{V(x, x^k)}}_{\le 0}\right)$$

$$f\left(\frac{1}{k}\sum \ x^{k+1}\right) - f(x) \le \frac{V(x, x^0)}{\gamma k}$$

$$f(\bar{x}^k) - f(x^*) \le \frac{\underline{L V(x^*, x^0)}}{k} \qquad \gamma = \frac{1}{L}$$

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

$\|\cdot\|_\infty \le \|\cdot\|_2$ $\quad \sqrt{d}$

$\|\cdot\|_2 \le \|\cdot\|_1$

$\|\nabla f(x) - \nabla f(y)\|_q \le L\|x-y\|_p$ $\qquad \frac{1}{p} + \frac{1}{q} = 1$ $\qquad p = 1 \Rightarrow q = \infty$

$\|\cdot\|_\infty \le L\|\cdot\|_1$ $\qquad \|\cdot\|_2 \le L_2\|\cdot\|_2$

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 $\gamma \le \frac{1}{L}$. Тогда справедлива следующая оценка сходимости

$L_2 \gg L$

$a \le b$ $\qquad a \le L B$ $\qquad b \le L_2 A$

$\|\cdot\|_\infty \quad \|\cdot\|_2$

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

$\|x^* - x^0\|_2^2$

$A \le B$ $\qquad L \sim \frac{a}{B}$

$\|\cdot\|_2 \le \|\cdot\|_1$

GD с евкл. нормой

$\ominus \quad \frac{L_2 \cdot \frac{1}{2}\|x^0 - x^*\|_2^2}{K} \quad \oplus$ $\qquad L_2 \sim \frac{b}{A}$ $\qquad \oplus \quad \frac{L\, V(x^*, x^0)}{K} \quad \ominus$

$L \le L_2$

$V(x,y) \ge \frac{1}{2}\|x-y\|_2^2$

$V(x,y) \ge \frac{1}{2}\|x-y\|^2 \ge \frac{1}{2}\|x-y\|_2^2$

Сходимость зеркального спуска

- Результат очень похож на сходимость градиентного спуска и обобщает его.

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

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

$$\|\nabla f(x) - f(y)\|_q \leq L\|x - y\|_p, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

- 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)?

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

$$\|\nabla f(x) - f(y)\|_q \leq L\|x - y\|_p, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

- 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.

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
$$\|\nabla f(x) - f(y)\|_q \leq L\|x - y\|_p, \quad \frac{1}{p} + \frac{1}{q} = 1.$$
- 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?

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
$$\|\nabla f(x) - f(y)\|_q \le L\|x - y\|_p, \quad \frac{1}{p} + \frac{1}{q} = 1.$$
- Если $1 \le p \le 2$, то $\|\cdot\|_2 \le \|\cdot\|_p, \quad \|\cdot\|_q \le \|\cdot\|_2$.
- **Вопрос:** что тогда можно сказать про отношение L и L_2(которую использовали ранее в евклидовом случае)? $L \le L_2$, а в хорошем случае L значительно меньше.
- **Вопрос:** с дивергенций тоже все хорошо? Там ситуация обратная для $1 \le p \le 2$, $V(x, y) \ge \frac{1}{2}\|x - y\|_p^2 \ge \frac{1}{2}\|x - y\|_2^2$

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
$$\|\nabla f(x) - f(y)\|_q \leq L\|x - y\|_p, \quad \frac{1}{p} + \frac{1}{q} = 1.$$
- 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

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

Пример зеркального спуска

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

$x_i \log(x_i)$

Найдем явный вид для $V(x,y) = \sum_{i=1}^{d} x_i \log\left(\frac{x_i}{y_i}\right)$ на $\triangle_d$. $= \{ x_i \geq 0,\ \sum x_i = 1 \}$

- Формальная запись задачи минимизации:

$d(x) - d(y) - \langle \nabla d(y); x - y \rangle$

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

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

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

$$L(x, \lambda, \nu) = \langle \gamma \nabla f(x^k); x \rangle + V(x, x^k) - \sum \lambda_i x_i + \nu\left(\sum x_i - 1\right)$$

$\sum x_i \log\left(\frac{x_i}{x_i^k}\right)$

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

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:

$$\begin{aligned} \min_{x \in \triangle_d} \ & \langle \gamma \nabla f(x^k), x \rangle + V(x, x^k) \\ \text{s.t.} \ & -x_i \leq 0 \\ & \sum_{i=1}^d x_i - 1 = 0 \end{aligned}$$

- 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

$$\cancel{\frac{x_i^k}{x_i}} \cdot \frac{1}{\cancel{x_i^k}}\,\cancel{x_i} + \log\left(\frac{x_i}{x_i^k}\right) + \underbrace{\gamma[\nabla f(x^k)]_i - \lambda_i + \nu}_{C} = 0$$

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

$$\frac{x_i}{x_i^k} = \exp(-1-C) \qquad x_i^* = x_i^k \exp(-1-C)$$

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

- Минимизируем по каждой x_i, чтобы получить двойственную:

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

- Двойственная задача:

$$\lambda_i^* = 0$$

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

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^*? $\lambda_i^* = 0$

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.

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 $\lambda_i^* = 0$):

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

$\nabla_x L(x, \underset{\overset{\|}{0}}{\lambda}, \nu) = 0$

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 $\lambda_i^* = 0$):

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

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

$$x_i^* = x_i^k \exp(-\gamma \nabla f_i) \cdot \exp(-\nu^* + 1)$$

- 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:

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

$x_i \geq 0$

$\sum x_i = 1$

$\sum x_i^* = 1 = \exp(1-\nu^*) \sum x_i^k \exp(-\gamma \dots)$

$\exp(1-\nu^*) = \frac{1}{\sum x_i^k \exp(-\gamma \dots)}$

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^*?

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^*? $\sum_{i=1}^d x_i^* = 1$

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.

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:

$\| \quad \|_1 \geq \| \quad \|_2 \geq \| \quad \|_\infty$

$\sqrt{d}$ $\quad$ $\sqrt{d}$

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

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.