Метод зеркального спуска

Семинарист: Данилова М.

Градиентный спуск

$$\min_{x\in\mathbb{R}^n} \quad f(x)$$

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k) = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\gamma_k}\|x - x^k\|_2^2 \right\}.$$

Метод проекции градиента

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

где $Q \subseteq \mathbb{R}^n$ — выпуклое замкнутое множество.

$$x^{k+1} = \pi_Q\left(x^k - \gamma_k \nabla f(x^k)\right) = \underset{x\in Q}{\operatorname{argmin}} \left\{ f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\gamma_k}\|x - x^k\|_2^2 \right\}.$$

Сопряженная норма

Определение 1. Пусть $\mathbb{R}^n$ – конечномерное евклидово пространство, $\|\cdot\|$ – произвольная норма в $\mathbb{R}^n$. Тогда сопряженной нормой для $\|\cdot\|$ называется норма $\|\cdot\|_*$, определенная как

$$\|y\|_* = \sup_{\|x\|\le 1} x^\top y.$$

Пример: Гёльдеровы нормы n–мерных векторов (ℓ_p–норма):

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} - \ell_p \text{ норма}, \quad p \in [1, \infty]$$

$$\|\cdot\|_q = (\|\cdot\|_p)_* - \ell_q \text{ норма}$$

$$\frac{1}{p} + \frac{1}{q} = 1$$

$$\|x\|_1 = \sum_{i=1}^n |x_i| \qquad \|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \qquad \|x\|_\infty = \max_i |x_i|$$

ℓ_1–норма сопряжена к ℓ_∞–норме и наоборот

ℓ_2–норма сопряжена к ℓ_2–норме

Метод зеркального спуска

Основные идеи:

- Для минимизации на множестве Q хотелось бы учесть его геометрию.
- Локальная геометрия функции f также может быть использована для построения более эффективного метода.
- Возможно, поможет изменение нормы.

Пример

Пусть Q является *единичным симплексом*:

$$\mathrm{Q} = \{x \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i = 1\}.$$

Векторы $x \in \mathrm{Q}$ можно интерпретировать как дискретные распределения вероятностей, поэтому Q также называют *вероятностым симплексом.*

Расстояния между элементами Q более естественно измерять с помощью метрик для вероятностных распределений. Например, с помощью дивергенции Кульбака-Лейблера:

$$\mathcal{KL}(x||y) = \sum_{i=1}^n x_i \ln \frac{x_i}{y_i}.$$

Новая модель функции

Рассмотрим евклидову норму $\|\cdot\|_2$ в $\mathbb{R}^n$.

- μ-сильно выпуклая относительно $\|\cdot\|$, если

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 .$$

- L-гладкой относительно $\|\cdot\|$, если

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2 .$$

Что будет, если заменить 2-норму на некоторую другую $\|\cdot\|$?

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\gamma_k} \|x - x_k\|^2 \right\}.$$

Или на другую величину, которая "хорошо согласуется"с $\|\cdot\|$?

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\gamma_k} V(x, x_k) \right\}.$$

Прокс-функцией

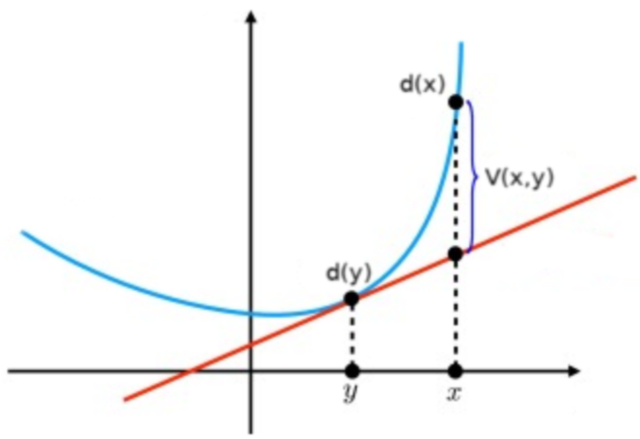
Определение 2. Прокс-функцией (distance generating function), связанной с нормой $\|\cdot\|$ для выпуклого замкнутого множества Q, назовем непрерывно дифференцируемую на $Q_0 \subseteq Q$ функцию $d(x)$, которая является 1-сильно выпуклой в норме $\|\cdot\|$, т.е.

$$d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2} \|y - x\|^2$$

Дивергенция Брэгмана

Определение 3. Дивергенцией Брэгмана (Bregman divergence), соответствующей прокс-функции $d(x) : Q \to \mathbb{R}$, назовем функцию $V(x, y) : Q \times Q_0 \to \mathbb{R}$, такую что

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



Свойства $V_d(x, y)$:

- Несимметричность: в общем случае $V_d(x, y) \neq V_d(y, x)$.
- $V_d(x, y)$ является сильно выпуклой по x.
- $V_d(x, y) \geq \frac{1}{2} \|y - x\|^2$ – следует из определения и 1-сильной выпуклости $d(x)$.
- (Three point equality) $V_d(z, x) + V_d(x, y) - V_d(z, y) = \langle \nabla d(y) - \nabla d(x), z - x \rangle$
  Аналогичное выполняется для $\|\cdot\|_2^2$:

$$\frac{1}{2} \|z - x\|_2^2 + \frac{1}{2} \|x - y\|_2^2 - \frac{1}{2} \|z - y\|_2^2 = \langle y - x, z - x \rangle .$$

- Евклидова прокс-функция $d(x) = \frac{1}{2} \|x\|_2^2$ порождает дивергенцию Брэгмана $V_d(x, y) = \frac{1}{2} \|x - y\|_2^2$.

- Пусть Q – единичный симплекс. Энтропийная прокс-функция $d(x) = \sum_{i=1}^n x_i \ln x_i$ порождает дивергенцию Брэгмана $V_d(x, y) = \sum_{i=1}^n x_i \ln \frac{x_i}{y_i}$, равную $\mathcal{KL}$-дивергенции между x и y.

Метод зеркального спуска

В шаге градиентного спуска заменим $\frac{1}{2}\left\|x - x^k\right\|_2^2$ на $V_d(x, x^k)$.

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathrm{Q}} \left[\underbrace{\left\langle \nabla f(x^k), x - x^k \right\rangle}_{\text{линейное приближение}} + \underbrace{\frac{1}{\gamma_k} V_d(x, x^k)}_{\text{отвечает за проекцию на Q}} \right].$$

Оказывается, как и в случае с методом проекции градиента, можно расщепить этот шаг на два:

$$\nabla d(y^k) = \nabla d(x^k) - \gamma_k \nabla f(x^k),$$
$$x^{k+1} = \operatorname*{argmin}_{x \in \mathrm{Q}} V_d(x, y^k).$$

Algorithm 1 Метод зеркального спуска

Require: Начальное приближение x^0, прокс-функция $d(x)$
for $k = 1, \dots, N$ **do**
Найти y^k из условия

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

Спроецировать y^k на Q относительно дивергенции Брэгмана:

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathrm{Q}} V_d(x, y^k)$$

end for

Скорость сходимости

Теорема 1. Пусть градиенты целевой функции f ограничены константой M, т.е. $\|\nabla f(x)\|_* \le M \quad \forall x \in \mathrm{Q}$. Кроме того, пусть число $R > 0$ такое, что $R^2 \geq 2 \inf_{x \in \mathrm{X}^*} V_d(x, x_0)$, где X^* – множество решений задачи $f(x) \to \min\limits_{x \in \mathrm{Q}}$.

Размер шага выбираем по правилу $h_k = \frac{\varepsilon}{M\left\|\nabla f(x^k)\right\|_*}$.

Тогда для всех $k \geq K = \frac{M^2 R^2}{\varepsilon^2}$ будет выполняться оценка

$$f(\bar{x}^k) - f^* \leq \varepsilon.$$

Константы M и R зависят от нормы $\|\cdot\|$. Хороший выбор нормы позволит уменьшить MR, и, следовательно, число итераций K.

Что значит "зеркальный"?

Градиент $\nabla d(x)$ задает отображение из Q_0 с нормой $\|\cdot\|$ (прямого пространства) в $\mathbb{R}^n$ с нормой $\|\cdot\|_*$ (двойственное пространство).

1. Точка x^k преобразуется в $\nabla d(x^k)$, лежащую в двойственном пространстве.
2. В двойственном пространстве выполняется градиентный шаг, и получается точка $\nabla d(y^k)$.
3. Точка $\nabla d(y^k)$ отображается в прямое пространство, и получается x^{k+1}. Это происходит с помощью проектирования относительно дивергенции Брэгмана.

Таким образом, градиентный спуск происходит в двойственном пространстве, а последовательность $\{x^k\}_{k=1}^N$ в прямом пространстве является его "отражением".

Пример

Рассмотрим задачу $f(x) \to \min\limits_{x \in \mathrm{Q}}$, где Q – единичный симплекс в $\mathbb{R}^n$.

Фиксируем норму $\|\cdot\|_1$ и возьмем энтропийную прокс-функцию

$$d(x) = \sum_{i=1}^n x_i \ln x_i.$$

Эта прокс-функция является 1-сильно выпуклой в $\|\cdot\|_1$. Ей соответствует дивергенция Брэгмана

$$V_d(x, y) = \sum_{i=1}^n x_i \ln \frac{x_i}{y_i}.$$

Шаг зеркального спуска принимает вид

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathrm{Q}} \left[\left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{\gamma_k} \sum_{i=1}^n x_i \ln \frac{x_i}{x_i^k} \right].$$

Эта задача имеет решение в явном виде

$$x^{k+1} = \frac{x^k \exp(-\gamma_k \nabla f(x^k))}{\sum\limits_{i=1}^n x_i^k \exp(-\gamma_k \frac{\partial f}{\partial x_i}(x^k))}.$$

Здесь x_i^k – i-ая компонента вектора x^k, а $\exp(-\gamma_k \nabla f(x^k))$ берется покомпонентно.

Анализ скорости сходимости

Сравним скорости сходимости зеркального спуска в нормах $\|\cdot\|_1$ и $\|\cdot\|_2$. В качестве начальной точки возьмем $x^0 = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ и оценим $R^2 = V_d(x^*, x^0)$ для каждого из случаев.

1. Норма $\|\cdot\|_1$, сопряженная норма $\|\cdot\|_\infty$: $R_1^2 = 2 \ln n$, число итераций $N_1 = O\left(\frac{M_\infty^2 \ln n}{\varepsilon^2}\right)$.

2. Норма $\|\cdot\|_2$, сопряженная норма $\|\cdot\|_2$: $R_2^2 = 1 - \frac{1}{n}$, число итераций $N_2 = O\left(\frac{M_2^2}{\varepsilon^2}\right)$.

Здесь M_∞ и M_2 – верхние оценки $\|\nabla f(x)\|_\infty$ и $\|\nabla f(x)\|_2$ соответственно.
Для любого вектора $x \in \mathbb{R}^n$ выполняется $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$

$$N_1 = O\left(\frac{M_\infty^2 \ln n}{\varepsilon^2}\right) \qquad \text{vs} \qquad N_2 = O\left(\frac{M_2^2}{\varepsilon^2}\right)$$

Для любого вектора $x \in \mathbb{R}^n$ выполняется $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$. Следовательно, $M_\infty \leq M_2 \leq \sqrt{n}M_\infty$, и

$$K_1 \leq O(K_2 \ln n) \leq O(nK_1)$$

- ЗС в $\|\cdot\|_1$ точно делает не более, чем в $O(\ln n)$ больше итераций по сравнению с ЗС в $\|\cdot\|_2$.
- Случай, когда $M_2 \approx \sqrt{n}M_\infty$, вполне возможен, если компоненты градиента $\nabla f(x)$ не сильно отличаются в точках множества Q.
- В последнем случае получим $K_1 \sim K_2 \frac{\ln n}{n}$, т.е. выигрыш по итерациям в $\frac{n}{\ln n}$ раз. Это существенно в пространствах большой размерности.

Выводы

- Хороший выбор нормы позволяет лучше учитывать геометрию допустимого множества или кривизну целевой функции.
- Аналог евклидового расстояния – дивергенция Брэгмана.
- Изменение нормы приводит к другому пониманию проектирования.
- В конечном итоге, можно получить выигрыш по количеству итераций, особенно в пространствах большой размерности.