

Субградиентный спуск. AdaGrad, Adam, DoG. Регуляризация ℓ_1 , ℓ_2 . Проксимальный оператор

Александр Безносиков

ИСП РАН

20 ноября 2024

Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла?

Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая?

Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая? Нет.
- Получается, что даже довольно простые выпуклые задачи могут быть негладким. До этого мы смотрели только на гладкие задачи.

Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая? Нет.
- Получается, что даже довольно простые выпуклые задачи могут быть негладким. До этого мы смотрели только на гладкие задачи.
- Будем рассматривать следующее предположение вместо гладкости (Липшицевости градиента):

Определение M -Липшецевой функции

Пусть дана функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Будем говорить, что она является M -Липшицева, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$|f(x) - f(y)| \leq M \|x - y\|_2.$$

Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая? Нет.
- Получается, что даже довольно простые выпуклые задачи могут быть негладким. До этого мы смотрели только на гладкие задачи.
- Будем рассматривать следующее предположение вместо гладкости (Липшицевости градиента):

Определение M -Липшецевой функции

Пусть дана функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Будем говорить, что она является M -Липшицева, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$|f(x) - f(y)| \leq M \|x - y\|_2.$$

Понятие (и все результаты далее) можно перенести на некоторое ограниченное выпуклое множество \mathcal{X} . Связано это в том числе с тем, что не бывает сильно выпуклых и Липшецевых на \mathbb{R}^d функций.

Вопрос: почему?

Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая? Нет.
- Получается, что даже довольно простые выпуклые задачи могут быть негладким. До этого мы смотрели только на гладкие задачи.
- Будем рассматривать следующее предположение вместо гладкости (Липшицевости градиента):

Определение M -Липшецевой функции

Пусть дана функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Будем говорить, что она является M -Липшицева, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$|f(x) - f(y)| \leq M \|x - y\|_2.$$

Понятие (и все результаты далее) можно перенести на некоторое ограниченное выпуклое множество \mathcal{X} . Связано это в том числе с тем, что не бывает сильно выпуклых и Липшецевых на \mathbb{R}^d функций.

Вопрос: почему? Линейный и квадратичный рост не сочетаются.

Субградиент и субдифференциал

Если функция не дифференцируема в точке, а значит градиента нет. Что может существовать вместо градиента?

Субградиент и субдифференциал

Если функция не дифференцируема в точке, а значит градиента нет. Что может существовать вместо градиента?

Субградиент и субдифференциал

Пусть дана выпуклая функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Вектор g будем называть субградиентом этой функции f в точке $x \in \mathbb{R}^d$, если для любого $y \in \mathbb{R}^d$ выполняется:

$$f(y) \geq f(x) + \langle g, y - x \rangle.$$

Множество $\partial f(x)$ всех субградиентов f в x будем называть субдифференциалом.

Условие оптимальности

Теорема (условие оптимальности)

x^* – минимум выпуклой функции f тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

Условие оптимальности

Теорема (условие оптимальности)

x^* – минимум выпуклой функции f тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

Доказательство:

\Leftarrow Если $0 \in \partial f(x^*)$, то по выпуклости и определению субградиента:
 $f(x) \geq f(x^*) + \langle 0, x - x^* \rangle = f(x^*)$. Доказано по определению глобального минимума.

Условие оптимальности

Теорема (условие оптимальности)

x^* – минимум выпуклой функции f тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

Доказательство:

\Leftarrow Если $0 \in \partial f(x^*)$, то по выпуклости и определению субградиента:
 $f(x) \geq f(x^*) + \langle 0, x - x^* \rangle = f(x^*)$. Доказано по определению глобального минимума.

\Rightarrow Если $f(x) \geq f(x^*)$ для любых $x \in \mathbb{R}^d$, то для вектора 0 выполнено
 $f(x) \geq f(x^*) + \langle 0, x - x^* \rangle$ для любого $x \in \mathbb{R}^d$. Доказано по определению субградиента.

Свойство M -Липшицевой функции

Лемма (свойство M -Липшицевой функции)

Пусть дана выпуклая функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Тогда функция f является M -Липшицевой тогда и только тогда, когда для любого $x \in \mathbb{R}^d$ и $g \in \partial f(x)$ имеем $\|g\|_2 \leq M$.

Субградиентный метод

- Рассматриваем задачу:

$$\min_{x \in \mathbb{R}^d} f(x),$$

где f выпуклая и M -Липшицева.

Субградиентный метод

- Рассматриваем задачу:

$$\min_{x \in \mathbb{R}^d} f(x),$$

где f выпуклая и M -Липшицева.

- Простая идея – вместо градиента использовать какой-то субградиент в текущей точке:

Алгоритм 2 Субградиентный метод

Вход: размеры шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $g^k \in \partial f(x^k)$
- 3: $x^{k+1} = x^k - \gamma g^k$
- 4: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^{K-1} x^k$

Доказательство сходимости

- Ничего сверхъестественного:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma g^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma \langle g^k, x^k - x^* \rangle + \gamma^2 \|g^k\|_2^2\end{aligned}$$

Доказательство сходимости

- Ничего сверхъестественного:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma g^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma \langle g^k, x^k - x^* \rangle + \gamma^2 \|g^k\|_2^2\end{aligned}$$

- Из M -Липшицевости f следует, что субградиенты ограничены:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma \langle g^k, x^k - x^* \rangle + \gamma^2 M^2$$

Доказательство сходимости

- Ничего сверхъестественного:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma g^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma \langle g^k, x^k - x^* \rangle + \gamma^2 \|g^k\|_2^2\end{aligned}$$

- Из M -Липшицевости f следует, что субградиенты ограничены:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma \langle g^k, x^k - x^* \rangle + \gamma^2 M^2$$

- Из выпуклости и определения субградиента:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma (f(x^k) - f(x^*)) + \gamma^2 M^2$$

Доказательство сходимости

- С предыдущего слайда:

$$2\gamma(f(x^k) - f(x^*)) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

Доказательство сходимости

- С предыдущего слайда:

$$2\gamma(f(x^k) - f(x^*)) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

- Суммируем по всем k и усредняем:

$$\frac{2\gamma}{K} \sum_{k=0}^{K-1} (f(x^k) - f(x^*)) \leq \frac{\|x^0 - x^*\|_2^2}{K} + \gamma^2 M^2$$

Доказательство сходимости

- С предыдущего слайда:

$$2\gamma(f(x^k) - f(x^*)) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

- Суммируем по всем k и усредняем:

$$\frac{2\gamma}{K} \sum_{k=0}^{K-1} (f(x^k) - f(x^*)) \leq \frac{\|x^0 - x^*\|_2^2}{K} + \gamma^2 M^2$$

- Откуда

$$\frac{1}{K} \sum_{k=0}^{K-1} (f(x^k) - f(x^*)) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

Доказательство сходимости

- С предыдущего слайда:

$$\frac{1}{K} \sum_{k=0}^{K-1} (f(x^k) - f(x^*)) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

- Гладкости нет, поэтому не получится доказать, что $f(x^k) \leq f(x^{k-1})$. Поэтому просто неравенство Йенсена для выпуклой функции:

$$f\left(\frac{1}{K} \sum_{k=0}^{K-1} x^k\right) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

Доказательство сходимости

- С предыдущего слайда:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

- Вопрос:** как подобрать шаг?

Доказательство сходимости

- С предыдущего слайда:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

- Вопрос:** как подобрать шаг? минимизировать правую часть по γ :
 $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$. Откуда

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right) - f(x^*) \leq \frac{M\|x^0 - x^*\|_2}{\sqrt{K}}$$

- Можно более практично: $\gamma_k \sim \frac{1}{\sqrt{k}}$.

Сходимость

Теорема сходимость субградиентного спуска для M -Липшицевых и выпуклых функций

Пусть задача безусловной оптимизации с M -Липшицевой, выпуклой целевой функцией f решается с помощью субградиентного спуска.

Тогда справедлива следующая оценка сходимости

$$f\left(\frac{1}{K} \sum_{k=0}^{K-1} x^k\right) - f(x^*) \leq \frac{M \|x^0 - x^*\|_2}{\sqrt{K}}$$

Более того, чтобы добиться точности ε по функции, необходимо

$$K = O\left(\frac{M^2 \|x^0 - x^*\|_2^2}{\varepsilon^2}\right) \text{ итераций.}$$

Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.

Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае?

Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.

Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум.

Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум. В общем случае результат для субградиентного метода является неулучшаемым для выпуклых и сильно-выпуклых задач, т.е. он оптимален.

Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум. В общем случае результат для субградиентного метода является неулучшаемым для выпуклых и сильно-выпуклых задач, т.е. он оптимален. **Вопрос:** а что в невыпуклом случае?

Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум. В общем случае результат для субградиентного метода является неулучшаемым для выпуклых и сильно-выпуклых задач, т.е. он оптимален. **Вопрос:** а что в невыпуклом случае? С этого мы начинали курс – лучше, чем полный перебор там ничего не придумать.

AdaGradNorm

- Для субградиентного метода был взят шаг $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$.
- Как уже было сказано, что можно взять k вместо K :
 $\gamma_k = \frac{\|x^0 - x^*\|_2}{M\sqrt{k}}$. **Вопрос:** как заменить его более практично – убрать M и $\|x^0 - x^*\|_2$, не теряя их физический смысл?

AdaGradNorm

- Для субградиентного метода был взят шаг $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$.
- Как уже было сказано, что можно взять k вместо K :
 $\gamma_k = \frac{\|x^0 - x^*\|_2}{M\sqrt{k}}$. **Вопрос:** как заменить его более практично – убрать M и $\|x^0 - x^*\|_2$, не теряя их физический смысл?
- M – ограничение нормы (суб)градиента, тогда можно использовать сам (суб)градиент в качестве этого ограничения, кроме этого $\|x^0 - x^*\|_2 \leq D$:

$$\gamma_k = \frac{D}{\sqrt{\sum_{t=0}^k \|g^t\|_2^2}}.$$

AdaGradNorm

- Получился метод AdaGradNorm. Ada – адаптивность под локальные свойства задачи (в данном случае локальные значения M).

Алгоритм 3 AdaGradNorm

Вход: $D > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов норм градиентов $G^0 = 0$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $g^k \in \partial f(x^k)$
- 3: Вычислить $G^{k+1} = G^k + \|g^k\|_2^2$
- 4: $x^{k+1} = x^k - \frac{D}{\sqrt{G^{k+1}}} g^k$
- 5: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^K x^k$

AdaGrad

- Пойдем дальше и сделаем адаптивность по каждой координате (индивидуальный шаг). Получится AdaGrad:

$$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^k (g_i^t)^2}}, \quad \text{где} \quad \|x_i - x_i^*\|_2 \leq D_i.$$

AdaGrad

Алгоритм 4 AdaGrad

Вход: $D_i > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов градиентов $G_i^0 = 0$, количество итераций K

1: **for** $k = 0, 1, \dots, K - 1$ **do**

2: Вычислить $g^k \in \partial f(x^k)$

3: Для каждой координаты: $G_i^{k+1} = G_i^k + (g_i^k)^2$

4: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1}}} g_i^k$

5: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^K x^k$

Доказательство сходимости AdaGrad

- Распишем шаг по каждой координате:

$$\begin{aligned}(x_i^{k+1} - x_i^*)^2 &= (x_i^k - \gamma_{k,i} g_i^k - x_i^*)^2 \\ &= (x_i^k - x_i^*)^2 - 2\gamma_{k,i} g_i^k (x_i^k - x_i^*) + \gamma_{k,i}^2 (g_i^k)^2.\end{aligned}$$

- Откуда:

$$g_i^k (x_i^k - x_i^*) = \frac{1}{2\gamma_{k,i}} (x_i^k - x_i^*)^2 - \frac{1}{2\gamma_{k,i}} (x_i^{k+1} - x_i^*)^2 + \frac{\gamma_{k,i}}{2} (g_i^k)^2.$$

Сходимость AdaGrad

Теорема сходимость AdaGrad для M -Липшицевых и выпуклых функций

Пусть задача оптимизации с M -Липшицевой, выпуклой целевой функцией f решается с помощью AdaGrad на ограниченном множестве. Тогда справедлива следующая оценка сходимости:

$$f\left(\frac{1}{K} \sum_{k=0}^{K-1} x^k\right) - f(x^*) \leq \frac{3M\tilde{D}}{2\sqrt{K}},$$

где $\tilde{D} = \sum_{i=1}^d D_i$.

Более того, чтобы добиться точности ε по функции, необходимо

$$K = O\left(\frac{9M^2\tilde{D}^2}{4\varepsilon^2}\right) \text{ итераций.}$$

RMSPprop

- Проблема AdaGrad, что старые градиенты в шаге могут быть уже не особо релевантны. **Вопрос:** как можно попробовать их «забывать»?

RMSPProp

- Проблема AdaGrad, что старые градиенты в шаге могут быть уже не особо релевантны. **Вопрос:** как можно попробовать их «забывать»?
- Может помочь техника импульса с $\beta_2 \in (0, 1)$ (вспомните, как она работала в случае тяжелого шарика):

$$G_i^{k+1} = \beta_2 G_i^k + (1 - \beta_2)(g_i^k)^2,$$

$$\gamma_{k,i} = \frac{D_i}{\sqrt{G_i^{k+1}}} g_i^k.$$

Получился метод RMSPProp.

RMSPProp

Алгоритм 5 RMSPProp

Вход: $D_i > 0$, моментум $\beta_2 \in (0,1)$, стартовая точка $x^0 \in \mathbb{R}^d$,
сглаженная сумма квадратов градиентов $G_i^0 = 0$, количество
итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $g^k \in \partial f(x^k)$
- 3: Для каждой координаты: $G_i^{k+1} = \beta_2 G_i^k + (1 - \beta_2)(g_i^k)^2$
- 4: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1}}} g_i^k$

5: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^K x^k$

Adam

- Почему бы не добавить и классический моментум вида тяжелого шарика для ускорения? Получится метод Adam:

тяжелый шарик: $v^{k+1} = \beta_1 v^k + (1 - \beta_1) g^k,$

RMSProp: $G_i^{k+1} = \beta_2 G_i^k + (1 - \beta_2)(g_i^k)^2,$

$$x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1}}} v_i^{k+1}.$$

Adam

- Почему бы не добавить и классический моментум вида тяжелого шарика для ускорения? Получится метод Adam:

тяжелый шарик: $v^{k+1} = \beta_1 v^k + (1 - \beta_1) g^k,$

RMSProp: $G_i^{k+1} = \beta_2 G_i^k + (1 - \beta_2)(g_i^k)^2,$

$$x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1}}} v_i^{k+1}.$$

- Можно еще чуть-чуть доработать – например, обезопасить себя от деления на 0 с помощью небольшой добавки $\epsilon \sim 10^{-6} - 10^{-8}$:

$$x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1} + \epsilon}} v_i^{k+1}$$

Adam

Алгоритм 6 Adam

Вход: $D_i > 0$, моменты $\beta_1 \in (0, 1)$ и $\beta_2 \in (0, 1)$, стартовая точка $x^0 \in \mathbb{R}^d$, сглаженная сумма квадратов градиентов $G_i^0 = 0$, сглаженная сумма градиентов $v^0 = 0$, добавка $\epsilon > 0$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $g^k \in \partial f(x^k)$
- 3: Вычислить $v^{k+1} = \beta_1 v^k + (1 - \beta_1) g^k$
- 4: Для каждой координаты: $G_i^{k+1} = \beta_2 G_i^k + (1 - \beta_2)(g_i^k)^2$
- 5: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1} + \epsilon}} v_i^{k+1}$

6: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^K x^k$

DoG

- Ситуация в общем случае даже ухудшилась по сравнению с AdaGrad: кроме шагов D_i , нужно подбирать еще моменты β_1, β_2 .

DoG

- Ситуация в общем случае даже ухудшилась по сравнению с AdaGrad: кроме шагов D_i , нужно подбирать еще моменты β_1, β_2 .
- Часто рекомендуют брать $\beta_1 = 0.9$, а $\beta_2 = 0.99$, и $D_i = D$ для всех i , но все равно нужно подбирать D .

DoG

- Ситуация в общем случае даже ухудшилась по сравнению с AdaGrad: кроме шагов D_i , нужно подбирать еще моментумы β_1, β_2 .
- Часто рекомендуют брать $\beta_1 = 0.9$, а $\beta_2 = 0.99$, и $D_i = D$ для всех i , но все равно нужно подбирать D .
- Проблема поиска D или D_i исконно пришла от того, что мы не знали $\|x^0 - x^*\|_2$. **Вопрос:** мы знаем x^0 , какая точка (в некотором смысле) является хорошим выбором в качестве приближения x^* ?

DoG

- Ситуация в общем случае даже ухудшилась по сравнению с AdaGrad: кроме шагов D_i , нужно подбирать еще моментумы β_1, β_2 .
- Часто рекомендуют брать $\beta_1 = 0.9$, а $\beta_2 = 0.99$, и $D_i = D$ для всех i , но все равно нужно подбирать D .
- Проблема поиска D или D_i исконно пришла от того, что мы не знали $\|x^0 - x^*\|_2$. **Вопрос:** мы знаем x^0 , какая точка (в некотором смысле) является хорошим выбором в качестве приближения x^* ?
- Можно предположить, что $x^k \approx x^*$ (метод приближается к решению):

$$\|x^0 - x^*\|_2 \rightarrow \|x^0 - x^k\|_2 \rightarrow d_k = \max_{t \in [k-1]} \|x^0 - x^t\|_2.$$

или

$$|x_i^0 - x_i^*| \rightarrow |x_i^0 - x_i^k| \rightarrow d_{k,i} = \max_{t \in [k-1]} |x_i^0 - x_i^t|.$$

DoG

- Имеем метод DoG:

$$\gamma_k = \frac{d_k}{\sqrt{\sum_{t=0}^k \|g^t\|_2^2}} \quad \text{или} \quad \gamma_{k,i} = \frac{d_{k,i}}{\sqrt{\sum_{t=0}^k (g_i^t)^2}}.$$

DoG

- Имеем метод DoG:

$$\gamma_k = \frac{d_k}{\sqrt{\sum_{t=0}^k \|g^t\|_2^2}} \quad \text{или} \quad \gamma_{k,i} = \frac{d_{k,i}}{\sqrt{\sum_{t=0}^k (g_i^t)^2}}.$$

- Почему DoG? Distance over Gradients:

$$d_k = \max_{t \in [k-1]} \left\| \sum_{\tau=0}^t \gamma_\tau \nabla f(x^\tau) \right\|_2.$$

DoG

Алгоритм 7 DoG v1

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов норм градиентов $G^0 = 0$, расстояние от точки старта $d_0 = 0$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $g^k \in \partial f(x^k)$
- 3: Вычислить $G^{k+1} = G^k + \|g^k\|_2^2$
- 4: Вычислить $d_{k+1} = \max\{d_k, \|x^0 - x^k\|\}$
- 5: $x^{k+1} = x^k - \frac{d_{k+1}}{\sqrt{G^{k+1}}} g^k$
- 6: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^K x^k$

DoG

Алгоритм 8 DoG v2

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов градиентов $G_i^0 = 0$, координатные расстояния от точки старта $d_{0,i}$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $g^k \in \partial f(x^k)$
- 3: Для каждой координаты: $G_i^{k+1} = G_i^k + (g_i^k)^2$
- 4: Вычислить $d_{k+1,i} = \max\{d_{k,i}; |x_i^0 - x_i^k|\}$
- 5: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{d_{k,i}}{\sqrt{G_i^{k+1}}} g_i^k$

6: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^K x^k$

Адаптивные методы: итог

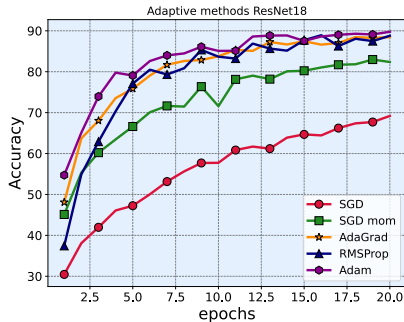
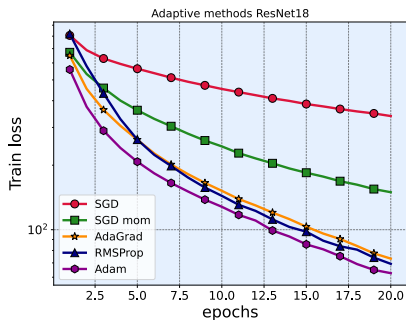
- Суть – подбирать шаг, исходя из локальных свойств задачи, что кажется более эффективным и удобным подходом.
- Adam и его модификации являются самым популярным методами решения задач оптимизации, лежащих в основе обучения нейронных сетей. Но в нем подбор параметров все еще нужен.

Адаптивные методы: итог

- Суть – подбирать шаг, исходя из локальных свойств задачи, что кажется более эффективным и удобным подходом.
- Adam и его модификации являются самым популярным методами решения задач оптимизации, лежащих в основе обучения нейронных сетей. Но в нем подбор параметров все еще нужен.
- Но эта проблема сейчас решена в так называемых parameter-free методах: кроме DoG можно обратить внимание на D-Adaptation/Prodigy.

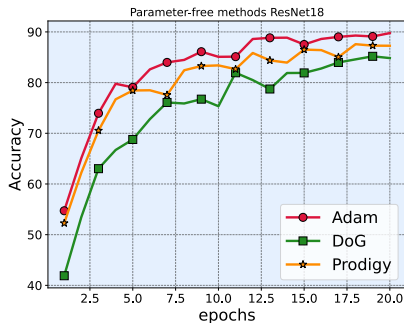
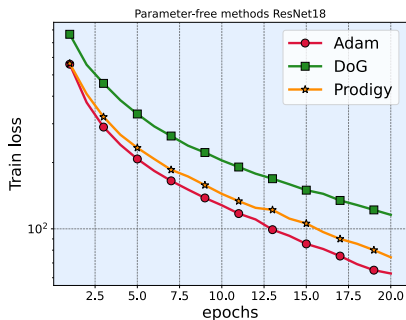
Адаптивные методы: сравнение на практике

- Для наглядности мы поставили эксперименты для задачи классификации на датасете CIFAR10 на модели RESNET18 (популярная модель глубокого обучения).



Parameter-free методы: сравнение на практике

- Аналогично сравним результаты работы DoG, рассмотренного ранее с одним из топовых беспараметрических методов – Prodigy.



Проксимальный оператор

- Поняли, что негладкие задачи «более сложные» по сравнению с гладкими задачами.
- Может быть получится «спрятать под ковер» отсутствие гладкости?

Проксимальный оператор

- Поняли, что негладкие задачи «более сложные» по сравнению с гладкими задачами.
- Может быть получится «спрятать под ковер» отсутствие гладкости?
- Такую возможность дает проксимальный оператор:

Определение проксимального оператора

Для функции $r : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ проксимальный оператор определяется следующим образом:

$$\text{prox}_r(x) = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|^2 \right).$$

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r . Если существует такая $\hat{x} \in \mathbb{R}^d$, что $r(x) < +\infty$. Тогда проксимальный оператор однозначно определен (т.е. всегда возвращает единственное уникальное значение).

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r . Если существует такая $\hat{x} \in \mathbb{R}^d$, что $r(\hat{x}) < +\infty$. Тогда проксимальный оператор однозначно определен (т.е. всегда возвращает единственное уникальное значение).

Доказательство: Проксимальный оператор возвращает минимум некоторой задачи оптимизации. Вопрос: что можно сказать про эту задачу?

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r . Если существует такая $\hat{x} \in \mathbb{R}^d$, что $r(x) < +\infty$. Тогда проксимальный оператор однозначно определен (т.е. всегда возвращает единственное уникальное значение).

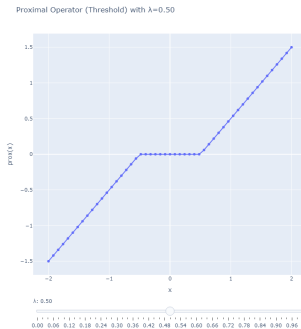
Доказательство: Проксимальный оператор возвращает минимум некоторой задачи оптимизации. Вопрос: что можно сказать про эту задачу? Она сильно выпуклая, а значит имеет строго один уникальный минимум (существование \hat{x} необходимо для того, чтобы $r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|^2$ где-то принимала конечное значение).

Примеры проксимального оператора

- $r(x) = \lambda \|x\|_1$, где $\lambda > 0$. Тогда

$$[\text{prox}_r(x)]_i = [|x_i| - \lambda]_+ \cdot \text{sign}(x_i)$$

Такой проксимальный оператор еще называют трешхолдом.



Примеры проксимального оператора

- $r(x) = \frac{\lambda}{2}\|x\|_2^2$, где $\lambda > 0$. Тогда

$$\text{prox}_r(x) = \frac{x}{1 + \lambda}.$$

Примеры проксимального оператора

- $r(x) = \frac{\lambda}{2}\|x\|_2^2$, где $\lambda > 0$. Тогда

$$\text{prox}_r(x) = \frac{x}{1 + \lambda}.$$

- $r(x) = \mathbb{I}_{\mathcal{X}}(x)$, где \mathcal{X} – выпуклое множество, и

$$\mathbb{I}_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$

Вопрос: чему равен prox ?

Примеры проксимального оператора

- $r(x) = \frac{\lambda}{2}\|x\|_2^2$, где $\lambda > 0$. Тогда

$$\text{prox}_r(x) = \frac{x}{1 + \lambda}.$$

- $r(x) = \mathbb{I}_{\mathcal{X}}(x)$, где \mathcal{X} – выпуклое множество, и

$$\mathbb{I}_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$

Вопрос: чему равен prox ?

$$\text{prox}_r(x) = \text{proj}_{\mathcal{X}}(x).$$

Примеры проксимального оператора

- $r(x) = \frac{\lambda}{2}\|x\|_2^2$, где $\lambda > 0$. Тогда

$$\text{prox}_r(x) = \frac{x}{1 + \lambda}.$$

- $r(x) = \mathbb{I}_{\mathcal{X}}(x)$, где \mathcal{X} – выпуклое множество, и

$$\mathbb{I}_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$

Вопрос: чему равен prox ?

$$\text{prox}_r(x) = \text{proj}_{\mathcal{X}}(x).$$

- И еще множество других примеров и их комбинаций.

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r . Тогда для любых $x, y \in \mathbb{R}^d$ следующие три условия являются эквивалентными:

- $\text{prox}_r(x) = y$,
- $x - y \in \partial r(y)$,
- $\langle x - y, z - y \rangle \leq r(z) - r(y)$ для любого $z \in \mathbb{R}^d$.

Доказательство

- Первое условие переписывается, как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

Доказательство

- Первое условие переписывается, как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

- Из условия оптимальности для выпуклой функции r это эквивалентно **вопрос**: чему?

Доказательство

- Первое условие переписывается, как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

- Из условия оптимальности для выпуклой функции r это эквивалентно **вопрос**: чему?

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

Доказательство

- Первое условие переписывается, как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

- Из условия оптимальности для выпуклой функции r это эквивалентно **вопрос**: чему?

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

Получили эквивалентность первого и второго условий.

Доказательство

- Первое условие переписывается, как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

- Из условия оптимальности для выпуклой функции r это эквивалентно **вопрос**: чему?

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

Получили эквивалентность первого и второго условий.

- Из определения субдифференциала, для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:

$$\langle g, z - y \rangle \leq r(z) - r(y).$$

Доказательство

- Первое условие переписывается, как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

- Из условия оптимальности для выпуклой функции r это эквивалентно **вопрос**: чему?

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

Получили эквивалентность первого и второго условий.

- Из определения субдифференциала, для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:

$$\langle g, z - y \rangle \leq r(z) - r(y).$$

В частности справедливо и для $g = x - y$.

Доказательство

- Первое условие переписывается, как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

- Из условия оптимальности для выпуклой функции r это эквивалентно **вопрос**: чему?

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

Получили эквивалентность первого и второго условий.

- Из определения субдифференциала, для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:

$$\langle g, z - y \rangle \leq r(z) - r(y).$$

В частности справедливо и для $g = x - y$. В обратную сторону тоже очевидно: для $g = x - y$ выполнено соотношение выше, значит $g \in \partial r(y)$.

Доказательство

- Первое условие переписывается, как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

- Из условия оптимальности для выпуклой функции r это эквивалентно **вопрос**: чему?

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

Получили эквивалентность первого и второго условий.

- Из определения субдифференциала, для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:

$$\langle g, z - y \rangle \leq r(z) - r(y).$$

В частности справедливо и для $g = x - y$. В обратную сторону тоже очевидно: для $g = x - y$ выполнено соотношение выше, значит $g \in \partial r(y)$. Лемма доказана.

Композитная задача

- Рассмотрим следующую задачу:

$$\min_{x \in \mathbb{R}^d} [f(x) + r(x)].$$

Композитная задача

- Рассмотрим следующую задачу:

$$\min_{x \in \mathbb{R}^d} [f(x) + r(x)].$$

- Такая задача называется композитной.
- Предположим, что f является L -гладкой выпуклой функцией, r выпуклой (необязательно гладкой, но) проксимально дружественной функцией.

Композитная задача

- Рассмотрим следующую задачу:

$$\min_{x \in \mathbb{R}^d} [f(x) + r(x)].$$

- Такая задача называется композитной.
- Предположим, что f является L -гладкой выпуклой функцией, r выпуклой (необязательно гладкой, но) проксимально дружественной функцией.
- Получается целевая функция состоит из гладкой и в общем случае негладкой части. Если $r \equiv 0$, то получаем гладкую задачу, которую умеем решать. Если $f \equiv 0$, то получаем негладкую задачу.

Проксимальный градиентный метод

Алгоритм 9 Проксимальный градиентный метод

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $\nabla f(x^k)$
- 3: $x^{k+1} = \text{prox}_{\gamma r}(x^k - \gamma \nabla f(x^k))$
- 4: **end for**

Выход: x^K

Проксимальный градиентный метод

Алгоритм 10 Проксимальный градиентный метод

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $\nabla f(x^k)$
- 3: $x^{k+1} = \text{prox}_{\gamma r}(x^k - \gamma \nabla f(x^k))$
- 4: **end for**

Выход: x^K

- Если r непрерывно дифференцируема, то условие оптимальности для подзадачи подсчета проксимального оператора записывается, как:

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

Проксимальный градиентный метод

Алгоритм 11 Проксимальный градиентный метод

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Вычислить $\nabla f(x^k)$
- 3: $x^{k+1} = \text{prox}_{\gamma r}(x^k - \gamma \nabla f(x^k))$
- 4: **end for**

Выход: x^K

- Если r непрерывно дифференцируема, то условие оптимальности для подзадачи подсчета проксимального оператора записывается, как:

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

- Откуда получаем так называемую неявную запись метода:

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

Сходимость проксимального метода

Теорема

Пусть задача композитной оптимизации с L -гладкой, μ -сильно выпуклой целевой функцией f и выпуклой (необязательно гладкой, но) проксимально дружественной функцией r решается с помощью проксимального градиентного спуска. Тогда при $y_k = \frac{1}{L}$ справедлива следующая оценка сходимости:

$$\|x^k - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K \|x^0 - x^*\|_2^2.$$

т.е. для достижения точности ε по норме аргумента ($\|x^K - x^*\| \leq \varepsilon$) необходимо

$$K = \left(\frac{L}{\mu} \log \frac{\|x^0 - x^*\|_2}{\varepsilon} \right) \text{ итераций.}$$