Методы оптимизации. Повторение лекций.

Корнилов Никита Максимович

Московский физико-технический институт

Декабрь 2025г

Скорости сходимости

- Сублинейная: $\|x^k - x^*\|_2 \le \dfrac{C}{k^\alpha},\ C > 0,\ \alpha > 0.$
- Линейная: $\|x^k - x^*\|_2 \le Cq^k,\ C > 0,\ 0 < q < 1.$ Или $\|x^{k+1} - x^*\| \le q\|x^k - x^*\|.$
- Сверхлинейная: $\|x^k - x^*\|_2 \le Cq^{k^p},\ C > 0,\ 0 < q < 1,\ p > 1.$
- Квадратичная: $\|x^k - x^*\|_2 \le Cq^{2^k},\ C > 0,\ 0 < q < 1.$ Или $\|x^{k+1} - x^*\| \le q\|x^k - x^*\|^2.$

Пусть f — μ-сильно выпуклая, тогда

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

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

Пусть f — L-гладкая, тогда

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

Пусть f — L-гладкая и μ-сильно выпуклая, тогда

$$LI \succeq \nabla^2 f(x) \succeq \mu I.$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

Оптимальность

Theorem (Необходимые условия локального минимума)

Пусть дана непрерывно дифференцируемая функция $f : \mathbb{R}^d \to \mathbb{R}$. *Если точка* x^* *является локальным минимумом, то* $\nabla f(x^*) = 0$.

Theorem (Достаточные условия глобального минимума)

Пусть дана выпуклая непрерывно дифференцируемая на $\mathbb{R}^d$ *функция* $f : \mathbb{R}^d \to \mathbb{R}$. *Если для некоторой точки* $x^* \in \mathbb{R}^d$ *верно, что* $\nabla f(x^*) = 0$, *то* x^* *– глобальный минимум* f *на всем* $\mathbb{R}^d$.

Theorem (Достаточные условия условного минимума)

Пусть дана выпуклая непрерывно дифференцируемая на $\mathbb{R}^d$ *функция* $f : \mathbb{R}^d \to \mathbb{R}$ *и выпуклое множество* $\mathcal{X}$. *Тогда* $x^* \in \mathcal{X}$ *– глобальный минимум* f *на* $\mathcal{X}$ *тогда и только тогда, когда для всех* $x \in \mathcal{X}$ *выполнено*

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0.$$

Начало безусловной оптимизации. Градиентный спуск

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

Задача: для выпуклых функций ищем глобальный минимум, а для невыпуклых – стационарную точку. Метод:

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

Интуиция: идти в сторону убывания функции или решение системы $\frac{dx_t}{dt} = -\nabla f(x_t) dt$ или минимизация ограничивающей параболы

$$x^{k+1} = \arg\min_x \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \|x - x^k\|_2^2 \right\}.$$

Шаг: оптимальное значение $\gamma_k = \frac{1}{L}$, для сильно выпуклых функций можно $\gamma_k = \frac{2}{(L+\mu)}$.

Сходимость GD с оптимальным шагом

- Сублинейная сходимость для L-гладких и выпуклых функций:

$$f(x^K) - f(x^*) \leq \frac{2L\|x^0 - x^*\|_2^2}{K}.$$

- Для невыпуклых функций сохраняется скорость сходимости по квадрату нормы градиента:

$$\|\nabla f(\hat{x}^K)\|_2^2 \leq \frac{2L(f(x^0) - f^*)}{K}.$$

- Линейная сходимость для L-гладких и μ-сильно выпуклых функций:

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

- Для функций, удовлетворяющих условию Поляка-Лоясевича, эта скорость сходимости сохраняется.

Выбор шага в GD

- Степенной шаг

$$\gamma_k := \frac{\gamma}{\delta + k^p}, \ \gamma > 0, \ \delta > 0, \ p > 0. \tag{1}$$

Наиболее часто применяются на практике $\gamma_k = \frac{\gamma}{k+1}$ и $\gamma_k = \frac{\gamma}{\sqrt{k+1}}$.

- Наискорейший спуск

$$\gamma_k^* = \arg\min_{\gamma_k > 0} f(x^k - \gamma_k \nabla f(x^k)). \tag{2}$$

Ищем аналитически или методами одномерной минимизации. Градиент в итоговой точке перпендикулярен направлению.

- Адаптивный подбор: на каждой итерации подбираем шаг $\gamma_{k+1} = \frac{1}{2L_{k+1}}$, увеличивая L_{k+1} в ρ раз, пока не выполнено условие $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L_{k+1}} \|\nabla f(x^k)\|_2^2$.

Выбор шага в GD II

- Шаг Поляка–Шора с параметром $\alpha > 0$:

$$\gamma_k^* := \frac{f(x^k) - f^*}{\alpha \|\nabla f(x^k)\|_2^2}. \tag{3}$$

Получен из минимизации

$$\|x^{k+1} - x^*\|_2^2 \le \|x^k - x^*\|_2^2 - 2\gamma_k(f(x^k) - f^*) + \gamma_k^2 \|\nabla f(x^k)\|_2^2.$$

Вместо f^* часто используют некоторую нижнюю оценку.

- Правила Армихо, Вольфа, Голдстейна и прочие.

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k) + \tau_k (x^k - x^{k-1}).$$

- Интуиция: использовать инерцию траектории $\tau_k \in [0.85, 0.95]$.
- Оптимальный для **квадратичных** сильно выпуклых задач: с $\gamma_k = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$, $\tau_k = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)$ справедлива оценка:

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

- Сходимость в общем случае: не лучше градиентного спуска в теории, но на практике может быть заметно лучше, особенно вблизи решения. Наблюдается волнообразная сходимость, так как идёт не по направлению убывания.
- Особенности: нужно хранить дополнительный вектор x^{k-1} и подбирать два параметра γ_k с τ_k.

Ускоренный метод Нестерова

$$y^{k+1} = x^k + \tau_k(x^k - x^{k-1}), \quad x^{k+1} = y^{k+1} - \gamma_k \nabla f(y^{k+1}).$$

Интуиция: смотрим по инерции в будущее и оттуда делаем градиентный шаг с $\tau_k \in [0.85, 0.95]$.

Сходимость для L-гладких и μ-сильно выпуклых функций с $\gamma_k = \frac{1}{L}$, $\tau_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$:

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

Сходимость для L-гладких и выпуклых функций с $\gamma_k = \frac{1}{L}$, $\tau_k = \frac{k}{k+3}$:

$$f(x^K) - f(x^*) \leq \frac{4L\|x^0 - x^*\|_2^2}{(K+2)^2}.$$

Особенность: нужно хранить два доп вектора, подбирать два параметра γ_k и τ_k, оценки совпадают с нижними оценками.

Будем рассматривать следующий класс алгоритмов:

$$x^{k+1} \in x^0 + \mathrm{span}\{\nabla f(x^0), \ldots, \nabla f(x^k)\}. \tag{4}$$

Пусть задача безусловной оптимизации с L-гладкой, μ-сильно выпуклой функцией f решается методом первого порядка. Тогда для достижения точности ε по аргументу ($\|x^K - x^*\|_2 \leq \varepsilon$) потребуется

$$\Omega\left(\sqrt{\frac{L}{\mu}} \log \frac{\|x^0 - x^*\|_2}{\varepsilon}\right) \quad \text{оракульных вызовов.} \tag{5}$$

Пример плохой функции

Зафиксируем K — кол-во итераций метода и построим

$$f(x) = \frac{L-\mu}{8}\langle x, Ax\rangle - \frac{L-\mu}{4}\langle e_1, x\rangle + \frac{\mu}{2}\|x\|_2^2,$$

где матрица размерности $d = 2K$ задана следующим образом:

$$A = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 2 \end{pmatrix}.$$

Оптимум равен $x_i^* = -\frac{q^{2d+2}}{1-q^{2d+2}}\frac{1}{q^i} + \frac{1}{1-q^{2d+2}}q^i$ с $q = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.
Если начальная точка $x^0 = (0, 0, \dots, 0)^\top$, то за один шаг любого метода сможем заполнить только **одну** следующую координату для x^k.

Метод сопряженных градиентов

Используется для решения СЛАУ: $Ax = b, A \succ 0$.
Идея: разложить решение $Ax^* = b$ в базис из сопряженных относительно A направлений ($p_i^\top A p_j = 0, i \neq j$), т.е., $x^* = \sum_{k=0}^{d-1} \alpha_k p_k$, восстанавливая на k-ой итерации p_k и α_k.
1) По индукции доказываем, что следующее направление p_k сопряженно со всеми предыдущими

$$r_k = Ax^k - b = \nabla f(x^k), \quad p_k = -r_k + \beta_k p_{k-1}, \quad p_{-1} = 0.$$

Сопряженность p_{k-1} и p_k:

$$\beta_k = \frac{p_{k-1}^T A r_k}{p_{k-1}^T A p_{k-1}}.$$

2) По сопряженности считаем коэффициент α_k из представления x^*:

$$\alpha_k = \frac{p_k^T b}{p_k^T A p_k}, \quad x^{k+1} = x^k + \alpha_k p_k.$$

Метод сопряженных градиентов: особенности

- С квадратной положительно определенной матрицей размера d находит точное решение за не более чем d итераций (за число уникальных собственных значений).
- Сходимость по норме

$$\|x^k - x^*\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \|x^0 - x^*\|_A.$$

  Здесь $\|x\|_A^2 = x^T A x$ и $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.
- Обобщения для произвольных функций: $r_k = \nabla f(x^k)$, $\beta_{k+1} = \frac{\langle \nabla f(x^{k+1}), \nabla f(x^{k+1}) \rangle}{\langle \nabla f(x^k), \nabla f(x^k) \rangle}$ (Флетчер - Ривс) или $\beta_{k+1} = \frac{\langle \nabla f(x^{k+1}), \nabla f(x^{k+1}) - \nabla f(x^k) \rangle}{\langle \nabla f(x^k), \nabla f(x^k) \rangle}$ (Поляк - Рибьер), α_k - правило подбора шага. Полезно использовать рестарты.

Метод Ньютона

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k).$$

- Интуиция: минимизация приближения Тейлора второго порядка

$$x^{k+1} = \arg\min_{y \in \mathbb{R}^d} \left[f(x^k) + \langle \nabla f(x^k), y - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x^k), y - x^k \rangle \right]$$

- Сходимость для μ-сильно выпуклых функций с M-Липшицевым гессианом: Квадратичная, но локальная скорость

$$\|x^{k+1} - x^*\|_2 \leq \frac{M}{2\mu} \|x^k - x^*\|_2^2.$$

- Особенности: дорогая итерация из подсчета и обращения гессиана за $O(d^3)$, вне области квадратичной сходимости может расходиться.

Модификации метода Ньютона

- Демпированный метод Ньютона: сходится сублинейно вне области квадратичной сходимости

$$x^{k+1} = x^k - \gamma_k \underbrace{\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)}_{=p_k}, \quad \gamma_k = \arg\min_{\gamma} f(x^k - \gamma p_k)$$

- Усечённый метод Ньютона: считать p_k, решая несколько итераций метода CG для $\left(\nabla^2 f(x^k)\right) \cdot p_k = \nabla f(x^k)$.

Кубический Ньютон

Идея: минимизировать разложение до 3 порядка с параметром M_k

$$x^{k+1} = \arg\min_{y \in \mathbb{R}^d} [\langle \nabla f(x^k), y - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(y - x^k), y - x^k \rangle + \frac{M_k}{6} \|x^k - y\|_2^3]$$

эквивалентна **выпуклой одномерной задаче**

$$\min_{r \in D} \left[\frac{1}{2} \langle (\nabla^2 f(x^k) + \frac{M_k r}{2} I_d)^{-1} \nabla f(x^k), \nabla f(x^k) \rangle + \frac{M_k}{12} r^3 \right],$$

где $D = \{r \in \mathbb{R}_+ | \nabla^2 f(x^k) + \frac{M_k r}{2} I_d \succ 0\}$.
Локальная квадратичная сходимость даже для невыпуклых функций, плюс сублинейная сходимость вне области квадратичной.

Идея: приблизительный, быстрый пересчёт обратного гессиана H_{k+1}.
1) Квазиньютоновское урав: $s^k = x^{k+1} - x^k$ и $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$:

$$s^k = H_{k+1} y^k.$$

Обычно $H_0 = I_d$.
2) Нужно доп условие:

SR-1:

$$H_{k+1} = H_k + \mu_k q^k (q^k)^\top,$$

где $\mu_k \in \mathbb{R}$ и $q^k \in \mathbb{R}^d$ нужно подобрать.
Итоговая формула:

$$H_{k+1} = H_k + \frac{(s^k - H_k y^k)(s^k - H_k y^k)^\top}{(s^k - H_k y^k)^\top y^k}. \tag{6}$$

Broyden

Запишем квазиньютоновское уравнение для матрицы B: $y^k = B_{k+1}s^k$.
Доп условие минимальности

$$B_{k+1} = \arg\min_{B \in \mathbb{R}^{d\times d}} \|B - B_k\|_F^2$$
$$\text{s.t.} \quad Bs^k = y^k.$$

Итоговое одноранговое несимметричное решение:

$$B_{k+1} = B_k + \frac{(y^k - B_k s^k)(s^k)^\top}{(s^k)^\top s^k}.$$

Аналитический подсчет обратной матрицы:

$$H_{k+1} = H_k + \frac{(s^k - H_k y^k)(s^k)^\top H_k}{(s^k)^\top H_k y^k}.$$

Квазиньютоновские методы

DPF:

$$B_{k+1} = \arg \min_{B \in \mathbb{R}^{d\times d}} \|B - B_k\|_W^2$$
$$s.t. \quad Bs^k = y^k$$
$$B^T = B,$$

где $\|A\|_W = \|W^{1/2}AW^{1/2}\|_F$ и $W = \int\limits_0^1 \nabla^2 f(x^k - \tau\gamma_k H_k \nabla f(x^k))d\tau$.
Итоговая формула:

$$B_{k+1} = (I_d - \frac{y^k(s^k)^\top}{(y^k)^\top s^k})B_k(I_d - \frac{s^k(y^k)^\top}{(y^k)^\top s^k}) + \frac{y^k(y^k)^\top}{(y^k)^\top s^k}.$$

Обращая ее по формуле Шермана-Моррисона-Вудбери, получаем:

$$H_{k+1} = H_k - \frac{H_k y^k (y^k)^\top H_k}{(y^k)^\top H_k y^k} + \frac{s^k(s^k)^\top}{(s^k)^\top y^k}.$$

Квазиньютоновские методы

BFGS:

$$H_{k+1} = \arg\min_{H \in \mathbb{R}^{d \times d}} \|H - H_k\|_W^2$$
$$s.t. \quad s^k = Hy^k$$
$$H^T = H$$

Итоговая формула:

$$H_{k+1} = (I_d - \frac{s^k (y^k)^\top}{(y^k)^\top s^k}) H_k (I_d - \frac{y^k (s^k)^\top}{(y^k)^\top s^k}) + \frac{s^k (s^k)^\top}{(y^k)^\top s^k}.$$

Локальная сверхлинейная сходимость: с близостью $\lambda_f(x) = \sqrt{\langle \nabla f(x), (\nabla^2 f(x))^{-1} \nabla f(x) \rangle}$ и начальным условием $\lambda_f(x^0) \leq \frac{\log \frac{3}{2}}{4} \frac{\mu^{5/2}}{LM}$ справедлива следующая оценка сходимости:

$$\lambda_f(x^K) \leq (\frac{11dL}{\mu K})^{K/2} \lambda_f(x^0).$$

L-BFGS:

Можно использовать только последние $m \approx 20$ шагов для апдейта матрицы H_{k+1}, нужно хранить лишь m пар векторов, а не матрицу

$$
\begin{aligned}
H_{k+1} = & \left(\left(V_{k-1}\right)^\top \ldots \left(V_{k-m}\right)^\top\right) H_k^0 \left(V_{k-m} \ldots V_{k-1}\right) \\
& + \rho^{k-m}\left(\left(V_{k-1}\right)^\top \ldots \left(V_{k-m+1}\right)^\top\right) s^{k-m}\left(s^{k-m}\right)^\top \left(V_{k-m+1} \ldots V_{k-1}\right) \\
& + \ldots \\
& + \rho^{k-1} s^{k-1}\left(s^{k-1}\right)^\top,
\end{aligned}
$$

где $V_k = I_d - \frac{y^k (s^k)^\top}{(y^k)^\top s^k}, \rho^k = \frac{1}{(y^k)^\top s^k}, H_k^0 = \frac{(s^{k-1})^\top y^{k-1}}{(y^{k-1})^\top y^{k-1}} I_d$.

Начало оптимизации на выпуклых множествах. Проекционный GD

Ставится задача минимизации на выпуклом множестве $\min_{x \in X} f(x)$.
Проекция на выпуклое замкнутое мн-во:

$$\Pi_X(x) := \arg\min_{y \in X} \frac{1}{2}\|x - y\|_2^2, \quad \|\Pi_X(x_1) - \Pi_X(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Для решения выполнено: $x^* = \Pi_X(x^* - \gamma \nabla f(x^*))$
GD с проекцией (все док-ва по 2 свойству аналогичны):

$$x^{k+1} = \Pi_X\left[x^k - \gamma_k \nabla f(x^k)\right].$$

Итерационная сходимость как у GD: $\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K \|x^0 - x^*\|_2^2$.
Примеры проекций с готовыми решениями:

- $X = \ell_2$-шар радиуса R с центром в 0: $\Pi_X(x) = \min\left\{1, \frac{R}{\|x\|_2}\right\} x$.
- $X = \left\{y \in \mathbb{R}^d \mid Ay = b\right\}$: $\Pi_X(x) = x - A^T(AA^T)^{-1}(Ax - b)$

Метод Франка Вульфа

$$
\begin{aligned}
s^k &= \arg\min_{s \in X} \langle s, \nabla f(x^k) \rangle, \\
\gamma_k = \frac{2}{k+2}, \quad x^{k+1} &= (1-\gamma_k)x^k + \gamma_k s^k.
\end{aligned}
$$

Сходимость итерационная:

$$
f(x^K) - f(x^*) \le \frac{2\max\{L\,\mathrm{diam}(X)^2, f(x^0) - f(x^*)\}}{K+2},
$$

где $\mathrm{diam}(X) := \max\limits_{x,y \in X} \|x-y\|_2$ – диаметр множества X.

Примеры подзадачи с готовыми решениями:

- $X = \ell_1$-шар радиуса R с центром в 0:
  $y^* = -R\mathrm{sign}(x_i)\mathbf{e_i},\, i = \arg\max_j |x_j|$,
- $X =$ Симплекс $\triangle = \left\{ y \in \mathbb{R}^d \mid y_i \ge 0, \sum_{i=1}^d y_i = R \right\}$:
  $y^* = R\mathbf{e_i}$, где $i = \arg\min_j x_j$.

Часто ведет к разряженным решениям из-за линейной задачи.

Зеркальный спуск

Ставится задача минимизации на множестве $\min_{x \in X} f(x)$.
Идея: использовать другие расстояния кроме Евклидового, чтобы подстроиться под геометрию и множества в задаче. Можно эффективно подстраиваться под меняющуюся геометрию (онлайн и распредленная оптмизация).

Обобщение сильной выпуклости и гладкости относительно нормы $\|\cdot\|$:

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

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

Дивергенцией Брэгмана (аналог метрики)

Пусть дана дифференцируемая 1-сильно выпуклая относительно нормы $\|\cdot\|$ на множестве X функция d (аналог нормы). Дивергенция Брэгмана $V(x,y) : X \times X \to \mathbb{R}_+$ это

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

Примеры:

$$\text{Шар с } \|\cdot\|_2 : \quad d(x) = \frac{1}{2}\|x\|_2^2 \Longrightarrow V(x,y) = \frac{1}{2}\|x-y\|_2^2.$$

$$\text{Энтропия с } \|\cdot\|_1 : \quad d(x) = \sum_{i=1}^{d} x_i \log x_i \Longrightarrow V(x,y) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i}.$$

Выполняется теорема Пифагора, как и для Евклидового расстояния (доказательства аналогичны GD). Также любая дивергенция ограничена снизу $V(x,y) \geq \frac{1}{2}\|x-y\|_2^2$.

Зеркальный спуск 2

Минимизация ограничивающей параболы, как в GD:

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

Эквивалентная запись

$$x^{k+1} = \Pi_X^{V(\cdot,\cdot)}\left[(\nabla d)^{-1}(\nabla d(x^k) - \gamma \nabla f(x^k))\right], (\nabla d)^{-1} = \nabla d^*.$$

Сходимость: на выпуклом множестве X с L-гладкой относительно нормы $\|\cdot\|$, выпуклой целевой функцией f и шагом $\gamma \leq \frac{1}{L}$

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

Константа гладкости отн. $\|\cdot\|$ может быть гораздо меньше, чем для $\|\cdot\|_2$ в GD с проекцией. При этом $V(x^*, x^0)$ возрастает не так сильно.
На единичном симплексе с энтропийным сетапом:

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

Негладкая задача

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

где f выпуклая и M-Липшицева.
Критерий: M-Липшецевость функции $\longleftrightarrow \|g\|_2 \leq M, \forall g \in \partial f(\cdot)$.
Субградиентный метод:

$$g^k \in \partial f(x^k), \quad x^{k+1} = x^k - \gamma g^k.$$

Можно добавить проекцию если минимизация на множестве.
Сходимость:

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

Оптимальная оценка (но медленнее GD для гладких функций) и маленький шаг $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$.

Идея: сделать подбор оптимального шага $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$ адаптивным.

Algorithm AdaGradNorm

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, $G^{-1} = 0$, параметры $\varepsilon \sim 10^{-8}$, $D > 0$, количество итераций K

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

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

3: $G^k = G^{k-1} + \|g^k\|_2^2$

4: $x^{k+1} = x^k - \frac{D}{\sqrt{G^k + \varepsilon}} g^k$

5: **end for**

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

DoG

Идея: сделать подсчет расстояния от начальной точки до решения адаптивным.

Algorithm DoG (Distance over Gradients)

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, $G^{-1} = 0$, $d_{-1} > 0$, параметр $\varepsilon \sim 10^{-8}$, количество итераций K

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

  Вычислить $g^k \in \partial f(x^k)$

  $G^k = G^{k-1} + \|g^k\|_2^2$

  $d_k = \max(d_{k-1}, \|x^k - x^0\|_2)$

  $x^{k+1} = x^k - \frac{d_k}{\sqrt{G^k + \varepsilon}} g^k$

end for

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

Следующий шаг — учесть неоднородность по координатам.

Algorithm AdaGrad

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, $G_i^{-1} = 0$, параметры $\varepsilon \sim 10^{-8}$, $D_i > 0$, количество итераций K

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

  Вычислить $g^k \in \partial f(x^k)$

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

  $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^k + \varepsilon}} g_i^k$

end for

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

Для M-Липшецевой и выпуклой функции верна оценка:

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

где $\tilde{D} = \sum_{i=1}^d D_i$, $|x_i^k - x_i^*| \leq D_i, \forall i = \overline{1, d}, k = \overline{0, K-1}$.

RMSProp

У AdaGrad знаменатель монотонно растёт, и шаг со временем может становиться слишком малым. Чтобы избежать затухания, используют экспоненциальное скользящее среднее квадратов градиентов.

Algorithm RMSProp

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, $G_i^{-1} = 0$, параметры $\varepsilon \sim 10^{-8}$, $D_i > 0$, $\beta \in [0, 1]$, количество итераций K

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

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

3: $G_i^k = \beta G_i^{k-1} + (1-\beta)(g_i^k)^2$

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

5: **end for**

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

Adam

Алгоритм Adam объединяет идеи RMSProp и момента: два моментума $\beta_1, \beta_2 \in [0, 1]$, обычно выбираются 0.9 и 0.999 соответственно.

Algorithm Adam

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, $G_i^{-1} = 0$, $v^{-1} = 0$, параметры $\beta_1, \beta_2 \in [0, 1]$, $\varepsilon \sim 10^{-8}$, $D_i > 0$, количество итераций K

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

  Вычислить $g^k \in \partial f(x^k)$

  $v^k = \beta_1 v^{k-1} + (1 - \beta_1) g^k$

  $\hat{v}^k = v^k / (1 - \beta_1^{k+1})$

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

  $\widehat{G}^k = G^k / (1 - \beta_2^{k+1})$

  $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{\widehat{G}_i^k + \varepsilon}} \hat{v}_i^k$

end for

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

AdamW

В классическом Adam если применяется ℓ_2-регуляризация, то градиент масштабируется адаптивным шагом. AdamW отделяет регуляризацию.

Algorithm AdamW

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, $G_i^{-1} = 0$, $v^{-1} = 0$, параметры $\beta_1, \beta_2 \in [0, 1]$, $\lambda > 0$, $\varepsilon \sim 10^{-8}$, $D_i > 0$, кол-во итераций K

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

  Вычислить $g^k \in \partial f(x^k)$

  $v^k = \beta_1 v^{k-1} + (1-\beta_1) g^k$

  $\hat{v}^k = v^k / (1 - \beta_1^{k+1})$

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

  $\widehat{G}^k = G^k / (1 - \beta_2^{k+1})$

  $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{\widehat{G}_i^k + \varepsilon}} \hat{v}_i^k - \lambda D_i x_i^k$

end for

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

Проксимальный оператор (обобщение градиентного шага)

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

Примеры:

- $r(x) = \lambda \|x\|_1$, тогда $[\text{prox}_r(x)]_i = [|x_i| - \lambda]_+ \cdot \text{sign}(x_i)$,
- $r(x) = \frac{\lambda}{2}\|x\|_2^2$, тогда $\text{prox}_r(x) = \frac{x}{1+\lambda}$.

Свойства:

- $\text{prox}_r(x) = y \quad \longleftrightarrow \quad x - y \in \partial r(y)$.
- $\langle x - y, \text{prox}_r(x) - \text{prox}_r(y) \rangle \geq \|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2$,
- $\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$.

Композитная задача и проксимальный метод

$$\min_{x \in \mathbb{R}^d} [f(x) + r(x)] \text{ – композитная задача,}$$

где f является L-гладкой выпуклой функцией, r выпуклой (необязательно гладкой, но проксимально дружественной функцией).
Проксимальный метод (сначала GD шаг по f, потом по r):

$$x^{k+1} = \mathrm{prox}_{\gamma_k r}(x^k - \gamma_k \nabla f(x^k)).$$

Альтернативая запись:

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

Сходимость: Композитная задача с L-гладкой, μ-сильно выпуклой целевой функцией f и выпуклой r при $\gamma_k = \frac{1}{L}$:

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

Точка x^* — неподвижная точка для $\mathrm{prox}_{\gamma_k r}(\cdot - \gamma_k \nabla f(\cdot))$

Начало оптимизации с ограничениями. Штрафная функция

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & f(x), \\ \text{s.t.} \quad & f_i(x) \le 0, \quad i = 1, \dots n, \\ & h_j(x) = 0, \quad j = 1, \dots m. \end{aligned} \tag{7}$$

Решение обычно располагается на границе допустимого множества. Аугментация:

$$\min_{x\in\mathbb{R}^d} \left[f_\rho(x) = f(x) + \rho \cdot \frac{1}{2}\sum_{i=1}^{n} (f_i^+)^2(x) + \rho \cdot \frac{1}{2}\sum_{j=1}^{m} h_j^2(x) \right].$$

Пусть все функции непрерывны и $\{x \in \mathbb{R}^d \mid f(x) \le f(x^*)\}$ ограничено. Тогда для любого $e > 0$ существует $\rho(e) > 0$ такое, что множество решений штрафной задачи X_ρ^* для любых $\rho \ge \rho(e)$ содержится в

$$X_e^* = \{x \in \mathbb{R}^d \mid \exists x^* \text{ решение задачи } (7) \; : \; \|x - x^*\|_2 \le e\}.$$

Метод штрафных функций

Алгоритм:

1. решить безусловную задачу $\min_{x \in \mathbb{R}^d} f_\rho(x)$ для текущего ρ, можно приближенно, используя один или несколько шагов стандартных методов;
2. увеличить ρ (увеличить силу штрафов);
3. использовать предыдущее решение как начальную точку;

Особенности:

- Условная задача превращена в безусловную.
- Подходим из вне допустимого множества к границе с решением. Увеличение ρ приближает к исходной задаче.
- При большом ρ наблюдается нарушение ограничений, что подходит не для всех задач.
- Увеличение ρ влечет за собой увеличение обусловленности задачи (константа Липшица градиента будет сильно расти). А значит задачу будет сложнее решать.

Двойственный подъем

Рассмотрим задачу с ограничениями

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & f(x) \\ \text{s.t.} \quad & Ax = b. \end{aligned}$$

Градиентный подъем для максимизации двойственной функции g:

$$\lambda^{k+1} = \lambda^k + \gamma_k \nabla_{\lambda_k} \left(\inf_{x\in\mathbb{R}^d} [f(x) + \lambda_k^\top (Ax - b)] \right).$$

Перепишем иначе (Теорема об огибающей):

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in\mathbb{R}^d} [f(x) + \lambda_k^\top (Ax - b)] = \arg\min_{x\in\mathbb{R}^d} L(x, \lambda^k), \\ \lambda^{k+1} &= \lambda^k + \gamma_k \nabla_{\lambda_k} (f(x^{k+1}) + \lambda_k^\top (Ax^{k+1} - b)) = \lambda^k + \gamma_k (Ax^{k+1} - b). \end{aligned}$$

В цикле решаем задачу для x^{k+1}, можно использовать один или несколько шагов стандартных методов, потом один шаг для λ^{k+1}.

Augmented Lagrangian, Hestenes-Powell

Рассмотрим эквивалентную задачу с аугментацией

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f(x) + \rho\|Ax - b\|_2^2 \\ \text{s.t.} \quad & Ax = b. \end{aligned}$$

Лагранжиан зависит от параметра и сильно выпуклый по x:

$$L_\rho(x, \lambda) = f(x) + \lambda_k^\top (Ax - b) + \rho\|Ax - b\|_2^2.$$

Двойственный подъем

$$\begin{aligned} x^{k+1} &= \arg\min_{x \in \mathbb{R}^d} L_\rho(x, \lambda^k), \\ \lambda^{k+1} &= \lambda^k + {\color{blue}\rho}(Ax^{k+1} - b). \end{aligned}$$

Особенности: надо подбирать только параметр ρ, сходится более стабильно.

ADMM

$$\min_{x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}} \quad f(x) + g(y),$$
$$\text{s.t.} \quad Ax + By = c,$$

где $A \in \mathbb{R}^{n \times d_x}$, $B \in \mathbb{R}^{n \times d_y}$, $c \in \mathbb{R}^n$.

Регуляризация:

$$\min_x f(x) + \alpha\|x\|_1 \qquad \min_{x,y} \quad f(x) + \alpha\|y\|_1, \quad \text{s.t.} \quad x = y.$$

Легко считать проксимальный оператор для отдельной функции.

Распределенная оптимизация:

$$\min_x \sum_{i=1}^n f_i(x) \quad \Leftrightarrow \quad \min_{x_i, y} \sum_{i=1}^n f_i(x_i), \quad \text{s.t.} \; x_i = y, i \in \overline{1, n}.$$

Каждую переменную x_i и функцию f_i можно оптимизировать отдельно.

ADMM

Аугментация

$$
\begin{aligned}
\min_{x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}} \quad & f(x) + g(y) + \frac{\rho}{2}\|Ax + By - c\|_2^2, \\
\text{s.t.} \quad & Ax + By = c.
\end{aligned}
$$

Минимум задачи тот же самый и не зависит от ρ. Лагранжиан:

$$
L_\rho(x, y, \lambda) = f(x) + g(y) + \lambda^\top (Ax + By - c) + \frac{\rho}{2}\|Ax + By - c\|_2^2.
$$

ADMM

Двойственный подъем:

$$
\begin{aligned}
x^{k+1} &= \arg\min_{x\in\mathbb{R}^{d_x}} L_\rho(x, y^k, \lambda^k),\\
y^{k+1} &= \arg\min_{y\in\mathbb{R}^{d_y}} L_\rho(x^{k+1}, y, \lambda^k)\\
\lambda^{k+1} &= \lambda^k + \rho\left(Ax^{k+1} + By^{k+1} - c\right),
\end{aligned}
$$

Вернуть $\frac{1}{K}\sum_{k=1}^K x^k, \frac{1}{K}\sum_{k=1}^K y^k, \frac{1}{K}\sum_{k=1}^K \lambda^k$.

Особенности: Вместо шага подбираем только ρ. Причем ρ не надо устремлять к бесконечности как в других методах. Позволяет разделять вычисления (распределенная оптимизация) и добавлять проксимальные операторы (l_1 регуляризации).

ADMM: Сходимость по зазору двойственности

Theorem

Если функции f и g являются выпуклыми и дружественными с точки зрения вычислений arg min, *то ADMM имеет следующую оценку сходимости для любого* $x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$, $\lambda \in \mathbb{R}^n$

$$L_0\left(\frac{1}{K}\sum_{k=1}^{K} x^k, \frac{1}{K}\sum_{k=1}^{K} y^k, \lambda\right) - L_0\left(x, y, \frac{1}{K}\sum_{k=1}^{K} \lambda^k\right) \le \frac{1}{2K}\|z^0 - z\|_P^2,$$

где L_0 *– Лагранжиан без аугментации,* $P = \begin{pmatrix} \rho A^T A & 0 & -A^T \\ 0 & 0 & 0 \\ -A & 0 & \frac{1}{\rho} I \end{pmatrix}$,

$z^0 = \begin{pmatrix} x^0 \\ y^0 \\ \lambda^0 \end{pmatrix}$

Барьерная функция

$$\min_{x\in\mathbb{R}^d} \quad f(x),$$
$$\text{s.t.} \quad f_i(x) \le 0, \quad i = 1, \dots n. \tag{8}$$

и допустимое множество $G = \{x \in \mathbb{R}^d | f_i(x) \le 0,\ i = \overline{1, n}\}$.
Барьером будем называть функцию $F : \mathbb{R}^d \to \mathbb{R}$:

- F непрерывно дифференцируема на $\mathrm{int}\, G$;
- Для любой последовательности $\{x_i\} \in \mathrm{int}\, G$ такой, что $x_i \to x \in \partial G$ (граница множества G), выполнено $F(x_i) \to +\infty$.

Примеры:

- Барьер Кэррола: $F(x) = -\sum_{i=1}^n \frac{1}{f_i(x)}$;
- Логарифмический барьер: $F(x) = -\sum_{i=1}^n \ln(-f_i(x))$.

Барьерная задача:

$$\min_{x\in \mathrm{int}\, G} \left[F_\rho(x) = f(x) + \frac{1}{\rho} F(x) \right].$$

Сходимость: Для любого $e > 0$ существует $\rho(e) > 0$ такое, что множество решений барьерной задачи X^*_ρ для любых $\rho \geq \rho(e)$ содержится в

$$X^*_e = \{x \in G \mid \exists x^* \text{ решение задачи (8) } : \|x - x^*\|_2 \leq e\}.$$

Алгоритм:

1. Увеличить $\rho_k > \rho_{k-1}$ (уменьшить силу барьера)
2. С помощью некоторого метода решить безусловную задачу $\min_{x \in \mathbb{R}^d} F_{\rho_k}(x)$ со стартовой точкой x_k. Можно приближенно, используя один или несколько шагов стандартных методов. Гарантировать, что выход метода x_{k+1} будет близок к реальному решению $x^*(\rho_k)$.

Всегда находимся внутри допустимого множества и подходим к границе с решением!

Самосогласованный барьер

Ограничивает темпы изменения кривизны и роста функции при приближении к границе. Позволяет четко оценить необходимые точность решения подзадачи и увеличение ρ.

Definition

Функция $B : \text{int } G \to \mathbb{R}$ является ν–самосогласованным барьером ($\nu \geq 1$) на множестве int G, если:

- B — самосогласованна на int G: для любых $x \in G$ и $h \in \mathbb{R}^d$:
$$\left| \frac{d^3}{dt^3} f(x + th) \right| \leq 2[\langle h, \nabla^2 f(x) h \rangle]^{\frac{3}{2}}.$$
- Для любой последовательности $\{x_i\} \subset \text{int } G$, сходящейся к $x \in \partial G$, выполнено $B(x_i) \to +\infty$.
- Для любых $x \in \text{int } G$ и $h \in \mathbb{R}^d$ выполнено условие:
$$|\langle h, \nabla B(x) \rangle| \leq \sqrt{\nu} \sqrt{\langle h, \nabla^2 B(x) h \rangle}.$$

Примеры самосогласованных барьеров

- Предложим n–самосогласованный барьер для множества выпуклых ограничений неравенств G с условием Слейтера: $B(x) = -\sum_{i=1}^{n} \log(-f_i(x))$.
- Для конуса $\{(t, x) \in \mathbb{R} \times \mathbb{R}^d | t \geq \|x\|_2\}$ подойдет 2–самосогласованный барьер $B(t, x) = -\log(t^2 - \langle x, x \rangle)$.
- Для матриц из $\mathbb{S}_+^d$ подойдет d–самосогласованный барьер $B(X) = -\log\det(X)$.

Свойства:

- Линейная комбинация ν_i–самосогласованных барьеров $B_i : \text{int } G_i \to \mathbb{R}$ с коэффициентами $\alpha_i \geq 1$ на $\cap \text{int } G_i$ является $\sum_{i=1}^{n} \alpha_i \nu_i$–самосогласованным барьером.
- Композиция ν–самосогласованного барьера и аффинного преобразования является ν–самосогласованным барьером.

Метод внутренней точки + самосогласованный барьер

Общая задача:

$$\min_{x \in \mathbb{R}^d} \quad f(x) \qquad \text{s.t.} \quad f_i(x) \le 0.$$

$$\min_{t \in \mathbb{R}, x \in \mathbb{R}^d} \quad t \qquad \text{s.t.} \quad f_i(x) \le 0, \quad f(x) - t \le 0.$$

$$\min_{x \in \mathbb{R}^d} \langle c, x \rangle \qquad \text{s.t.} \quad f_i(x) \le 0.$$

Используем ν-самосоглассованный барьер B и объекты:

$$\Phi_\rho(x) = \rho\langle c, x\rangle + B(x), \lambda(\Phi_\rho, x) = \sqrt{\langle (\nabla^2 \Phi_\rho(x))^{-1} \nabla \Phi_\rho(x), \nabla \Phi_\rho(x)\rangle}.$$

Будем применять метод внутренней точки и решать подзадачу с помощью одного шага демпфированного метода Ньютона

Метод внутренней точки (демпфированный метод Ньютона)

Algorithm Метод внутренней точки (демпфированный метод Ньютона)

Вход: стартовая точка $x^0 \in G$ такая, что $\lambda(\Phi_{\rho_{-1}}, x^0) \leq e_1$, параметры $e_1, e_2 \in (0,1)$, стартовое значение параметра $\rho_{-1} > 0$, K

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

2: $\quad \rho_k = (1 + \frac{e_2}{\sqrt{\nu}})\rho_{k-1}$

3: $\quad x^{k+1} = x^k - \frac{1}{1+\lambda(\Phi_{\rho_k}, x^k)}[\nabla^2 \Phi_{\rho_k}(x^k)]^{-1} \nabla \Phi_{\rho_k}(x^k)$

4: **end for**

Выход: x^K

Возможно, понадобится больше одного шага метода Ньютона, но при правильном соотношении e_1 и e_2 достаточно ровно одного. Также начальное приближение должно быть достаточно близко.

Сходимость метода внутренней точки

С настройкой $e_1 = 0.05$, $e_2 = 0.08$ и ν–самосогласованным барьером метод с выпуклыми ограничениями неравенства достигает ε решения по функции $(f(x) - f^* \leq \varepsilon)$ за

$$K = \tilde{O}\left(\sqrt{\nu} \log \frac{\nu}{\varepsilon \rho_0}\right) \text{ итераций.}$$

• Линейная скорость по ε и полиномиальная зависимость от количества переменных и ограничений.
• На практике можно куда быстрее изменять ρ, и число итераций метода Ньютона не превышает $30 - 50$.

Седловая задача

• Функция $L(x, y) : X \times \Lambda \to \mathbb{R}$. Контекст: x и λ это игроки, а $L(x, y)$ – цена, которую платит игрок x игроку λ. Все пытаются получить наилучшие условия.
• Седловая точка или равновесие (x^*, λ^*) — это такая точка, что для любых значений $x \in X$, $\lambda \in \Lambda$: $L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda)$.
• Седловые точки существуют $\Leftrightarrow$ они решают сразу две задачи

$$\max_{\lambda \in \Lambda} \min_{x \in X} L(x, \lambda) = \min_{x \in X} \max_{\lambda \in \Lambda} L(x, \lambda).$$

В общем случае, $\max_{\lambda \in \Lambda} \min_{x \in X} L(x, \lambda) \leq \min_{x \in X} \max_{\lambda \in \Lambda} L(x, \lambda)$. Для существования седловых точек достаточно, чтобы X, Λ были выпуклые и компактные + $L(x, \lambda)$ непрерывна и выпукла-вогнута: выпукла по x (фикс λ) и вогнута по λ (фикс x).

Связь с оптимизацией

Рассмотрим задачу условной оптимизации:

$$\begin{aligned} \min_{x \in X} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = \overline{1, m} \\ & Ax = b. \end{aligned}$$

Лагранжиан для этой задачи строится следующим образом:
$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^\top (Ax - b)$.

• С условием Слейтера любой набор оптимальных переменных x^* и (λ^*, ν^*) является седловой точкой лагранжиана:

$$L(x, \lambda^*, \nu^*) \geq g(\lambda^*, \nu^*) = f_0(x^*) = L(x^*, \lambda^*, \nu^*) \geq L(x^*, \lambda, \nu).$$

Решаем седловую задачу, максимизируя двойственную функцию:
$\max_{\lambda \geq 0, \nu \in \mathbb{R}^d} \min_{x \in X} L(x, \lambda, \nu)$.

Седловая задача и Спуск-Подъем

$$\min_{x\in\mathbb{R}^d} \max_{\lambda\in\mathbb{R}^d} L(x,\lambda),$$

где L выпукла-вогнута и L-гладкая функция

$$\|\nabla_x L(x_1,\lambda_1) - \nabla_x L(x_2,\lambda_2)\|_2^2 \leq \frac{L^2}{2}(\|x_1 - x_2\|_2^2 + \|\lambda_1 - \lambda_2\|_2^2),$$

$$\|\nabla_\lambda L(x_1,\lambda_1) - \nabla_\lambda L(x_2,\lambda_2)\|_2^2 \leq \frac{L^2}{2}(\|x_1 - x_2\|_2^2 + \|\lambda_1 - \lambda_2\|_2^2).$$

Обобщения GD расходятся или сходятся неоптимально особенно с просто выпукло-вогнутыми функциями ($L(x,y) = xy$):
Градиентный спуск подъем (Gradient Descent Ascent, GDA)

$$x^{k+1} = x^k - \gamma\nabla_x L(x^k,\lambda^k), \quad \lambda^{k+1} = \lambda^k + \gamma\nabla_\lambda L(x^k,\lambda^k)$$

или поочередный (Alt–GDA)

$$x^{k+1} = x^k - \gamma\nabla_x L(x^k,\lambda^k), \quad \lambda^{k+1} = \lambda^k + \gamma\nabla_\lambda L(x^{k+1},\lambda^k).$$

Экстраградиент

Идея: посмотреть в будущее, взять оттуда градиент и сходить по нему.

Экстраградиент

$$\begin{aligned}
x^{k+1/2} &= x^k - \gamma \nabla_x L(x^k, \lambda^k) \\
\lambda^{k+1/2} &= \lambda^k + \gamma \nabla_\lambda L(x^k, \lambda^k) \\
x^{k+1} &= x^k - \gamma \nabla_x L(x^{k+1/2}, \lambda^{k+1/2}) \\
\lambda^{k+1} &= \lambda^k + \gamma \nabla_\lambda L(x^{k+1/2}, \lambda^{k+1/2})
\end{aligned}$$

Вернуть $\frac{1}{K}\sum_{k=0}^{K-1} x^{k+1/2}, \frac{1}{K}\sum_{k=0}^{K-1} \lambda^{k+1/2}$

Экстраградиент: сходимость

Пусть дана выпукло-вогнутая и L-гладкая функция $L(x, \lambda)$. Для любого $u \in \mathbb{R}^d \times \mathbb{R}^n$ и для любого $\gamma \leq \frac{1}{L}$:

$$\left(L\left(\frac{1}{K} \sum_{k=0}^{K-1} x^{k+1/2}, u_\lambda \right) - L\left(u_x, \frac{1}{K} \sum_{k=0}^{K-1} \lambda^{k+1/2} \right) \right) \leq \frac{\|z^0 - u\|_2^2}{2\gamma K}$$

Метрика для решения на компактах: $\max_\lambda L(x^k, \lambda) - \min_x L(x, \lambda^k)$. См. $L(x, \lambda) = (x-1)(\lambda+1)$ с $x^* = 1, \lambda = -1$.

- Можно добавить проекции и решать седловую задачу на множествах $X \neq \mathbb{R}^d$ и $\Lambda \neq \mathbb{R}^n$ (2 проекции + 2 оракула градиента).
- Можно получить линейную сходимость для сильно выпуклых–сильно вогнутых задач.
- Часто применяют для решения двойственной задачи.

Модификации

Новые обозначения: $z = (x, \lambda)^\top, F(z) = (\nabla_x L(x, \lambda), -\nabla_\lambda L(x, \lambda))^\top$.

- **Past Extra–Gradient:**

$$z^{k+1/2} = \Pi_Z[z^k - \gamma_k F(z^{k-1/2})], \quad z^{k+1} = \Pi_Z[z^k - \gamma_k F(z^{k+1/2})].$$

Всего 1 градиент + 2 проекции.

- **Reflected Gradient:** смотрим градиент из прошлого

$$z^{k+1/2} = z^k - (z^{k-1} - z^k), \quad z^{k+1} = \Pi_Z[z^k - \gamma_k F(z^{k+1/2})].$$

Всего 1 градиент + 1 проекция.

- **Forward–Backward–Forward:** Смотрим вперед и на 2ом шаге

$$z^{k+1/2} = \Pi_Z[z^k - \gamma_k F(z^k)], \quad z^{k+1} = z^{k+1/2} + \gamma_k F(z^k) - \gamma_k F(z^{k+1/2}).$$

Всего 2 градиента + 1 проекция.

Модификации

- **Optimistic Gradient:** смотрим градиент из прошлого

  $$z^{k+\frac{1}{2}} = \Pi_Z[z^k - \gamma_k F(z^{k-\frac{1}{2}})], z^{k+1} = z^{k+\frac{1}{2}} + \gamma_k F(z^{k-\frac{1}{2}}) - \gamma_k F(z^{k+\frac{1}{2}}).$$

  Всего 1 градиент + 1 проекция.

Все алгоритмы выше сходятся как $1/K$.

- **Экстраград с моментумом** $\tau_k > 0$: уменьшает вихрения

  $$z^{k+1/2} = \Pi_Z[z^k - \gamma_k F(z^k)], z^{k+1} = \Pi_Z[z^k - \gamma_k F(z^{k+1/2}) + \tau_k(z^k - z^{k-1})]$$

- **Alt–GDA с моментумом** $\tau_k < 0$: хорош на практике

  $$x^{k+1} = x^k - \gamma_k \nabla_x L(x^k, \lambda^k) + \tau_k(x^k - x^{k-1}),$$

  $$\lambda^{k+1} = \lambda^k + \gamma_k \nabla_\lambda L(x^{k+1}, \lambda^k) + \tau_k(\lambda^k - \lambda^{k-1}).$$

Прямо-двойственный метод (Chambolle–Pock Algorithm)

- Рассмотрим задачу минимизации с выпуклой и гладкой f:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f(x) \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

с Ланграгнжианом $L(x, \lambda) = f(x) - \lambda^\top (Ax - b)$.

- Для следующей задачи можно использовать сопряженную функцию:

$$\min_{x \in \mathbb{R}^d} f(x) + h(Ax) \Leftrightarrow \min_{x \in \mathbb{R}^d} \max_{\lambda} \{f(x) + \lambda^\top Ax - h^*(\lambda)\}.$$

- В общем случае Лагранжиан $L(x, \lambda) = f(x) - \lambda^\top Ax + g(\lambda)$ с выпуклой, замкнутой и гладкой регуляризацией g.

Прямо-двойственный метод (Chambolle–Pock Algorithm)

Algorithm Прямо–двойственный алгоритм

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad x^{k+1} = x^k - \gamma_k(\nabla f(x^k) - A^\top \lambda^k)$
3: $\quad \lambda^{k+1} = \lambda^k - \gamma_k(\nabla g(\lambda^k) + A(2x^{k+1} - x^k))$
4: **end for**
Выход: $\frac{1}{K}\sum_{k=1}^{K} x^k, \frac{1}{K}\sum_{k=1}^{K} \lambda^k$

Сходится один в один как экстраградиент. В отличие от ADMM и двойственного подъема, переменные x^k и λ^k обновляются вместе эффективно, как у экстраградиента. Можно встроить проксимальный оператор для негладких f, g и добавить ускорение.

Онлайн-постановка:

$$\min_{x \in \mathbb{R}^d} [f(x) := \mathbb{E}_{\xi \sim D}[f(x, \xi)]]$$

Оффлайн-постановка:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^{n} [f(x, \xi_i)] \right]$$

Предположения (несмещенность и оценка на дисперсию шума):

$$\mathbb{E}_\xi [\nabla f(x, \xi)] = \nabla f(x), \quad \mathbb{E}_\xi [\|\nabla f(x, \xi) - \nabla f(x)\|_2^2] \leq \sigma^2.$$

SGD:

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k, \xi^k).$$

SGD Сходимость

Задача безусловной стохастической оптимизации с L-гладкой, μ-сильно выпуклой функцией f решается с помощью SGD с $\gamma_k \leq \frac{1}{L}$:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq (1 - \gamma_k \mu)\mathbb{E}\left[\|x^k - x^*\|^2\right] + \gamma_k^2 \sigma^2.$$

Постоянный шаг: сходимость до плато

$$\mathbb{E}\left[\|x^K - x^*\|^2\right] \leq (1 - \gamma\mu)^K \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\gamma\sigma^2}{\mu}.$$

Batching:

$$\nabla f(x^k, \xi^k) \to \frac{1}{b} \sum_{j \in S^k} \nabla f(x, \xi_j^k).$$

Уменьшение дисперсии с σ до $\sigma/\sqrt{b}$ в оценках выше.

Уменьшающийся шаг

- $\gamma_k = \frac{2}{\mu(k+1)}$:

$$\mathbb{E}[\|x^K - x^*\|_2^2] \leq \frac{\exp(8L^2/\mu^2)}{K^2}(\mathbb{E}[\|x^0 - x^*\|_2^2] + \frac{\sigma^2}{L^2}) + \frac{8\sigma^2 \log K}{\mu^2 K}.$$

- $\gamma_k = \frac{2}{\mu\sqrt{k+1}}$:

$$\mathbb{E}[\|x^K - x^*\|_2^2] \leq 2K^{16L^2/\mu^2}\exp(-\frac{\sqrt{K}}{2})(\mathbb{E}[\|x^0 - x^*\|_2^2]+\frac{\sigma^2}{L^2})+\frac{8\sigma^2}{\mu^2\sqrt{K}}.$$

- Сначала шаг постоянный, потом начинает уменьшаться.

Плюс: сублинейная сходимость до точного решения, минус: потеря быстрой сходимости в начале. Можно также применять батчинг.

SAGA

$$\min_x f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

Идея: уменьшаем с каждой итерацией дисперсию оценки градиента g^k до нуля, чтобы $g^k \to \nabla f(x^*) = 0$ при $x^k \to x^*$.

Алгоритм SAGA:

$$\begin{aligned}
&\text{Сгенерировать независимо } i_k\\
&g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n}\sum_{j=1}^{n} y_j^k\\
&y_i^{k+1} = \begin{cases} \nabla f_i(x^k), & \text{если } i = i_k\\ y_i^k, & \text{иначе} \end{cases}\\
&x^{k+1} = x^k - \gamma g^k
\end{aligned}$$

Несмещенная оценка: $\mathbb{E}_{i_k}\left[g^k | x^k\right] = \nabla f(x^k)$.

О сходимости SAGA

y_i^k запоминает с запаздыванием $\nabla f_i(x^k)$, а $\frac{1}{n}\sum_{j=1}^n y_j^k$ запоминает $\nabla f(x^k)$. При $x^k \to x^*$ имеем, что $y_j^k \to \nabla f_j(x^*)$, и $\frac{1}{n}\sum_{j=1}^n y_j^k \to \nabla f(x^*) = 0$, значит $g^k \to 0$.

Сходимость. Пусть задача безусловной стохастической оптимизации вида конечной суммы с L-гладкими, выпуклыми функциями f_i и μ-сильно выпуклой целевой функцией f решается с помощью SAGA с $\gamma \leq \frac{1}{6L}$. Тогда получается следующая **линейная** итерационная сложность:

$$\mathcal{O}\left(\left[n + \frac{L}{\mu}\right]\log\frac{1}{\varepsilon}\right).$$