Методы оптимизации. Повторение лекционных материалов.

Корнилов Никита Максимович

Московский физико-технический институт

Декабрь 2024г

Скорости сходимости

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

Свойства L-гладких и μ-сильно выпуклых функций

Пусть f — μ-сильно выпуклая, тогда

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

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

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

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

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

Оптимальность

Theorem

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

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

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

Шаг: оптимальное значение $\gamma_k = \frac{1}{2L}$.
Интуиция: решение системы $\frac{dx_t}{dt} = (x_t)dt$ или минимизация ограничивающей параболы

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

Сходимость для L-гладких функций:

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

Сходимость для L-гладких и μ-сильно выпуклых функций:

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

Метод тяжелого шарика

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

Интуиция: использовать инерцию траектории
Сходимость: не лучше градиентного спуска в теории

Ускоренный метод Нестерова

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

Сходимость 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 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 подбирать два параметра, совпадает с нижними оценками

Метод сопряженных градиентов

Используется для решения СЛАУ: $Ax = b, A \succ 0$.
Идея: разложить решение в базис сопряженных относительно A направлений $x^* = \sum_{i=0}^{d-1} \lambda_i p_i$, восстанавливая на каждой итерации p^k, λ_k.
По сопряженности:

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

По индукции доказываем, что след направление - сопряженное всем

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

Сопряженность p_{k-1}, p_k:

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

Метод сопряженных градиентов: особенности

- С квадратной положительно определенной матрицей размера d находит точное решение за не более чем d итераций.
- Сходимость с ошибкой

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

Здесь $\|x\|_A^2 = x^T A x$ и $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.

- Обобщения для произвольных функций: $r_k = \nabla f(x^k)$, α_k - правило подбора шага, $\beta_{k+1} = \frac{\langle \nabla f(x^{k+1}), \nabla f(x^{k+1})\rangle}{\langle \nabla f(x^k), \nabla f(x^k)\rangle}$ (Флетчер - Ривс) или $\beta_{k+1} = \frac{\langle \nabla f(x^{k+1}), \nabla f(x^{k+1}) - \nabla f(x^k)\rangle}{\langle \nabla f(x^k), \nabla f(x^k)\rangle}$ (Полак - Рибьер). Полезно использовать рестарты.

Метод Ньютона

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

Интуиция:

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

Сходимость: Квадратичная, но локальная скорость

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

Дорогая итерация $O(d^3)$, вне области сходимости может расходиться, но демпинг $\gamma_k \neq 1$ помогает

Квазиньютоновские методы (Итерация за $O(d^2)$)

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

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

Симметричность $H_{k+1} = H_{k+1}^\top$.

SR-1:

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

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

BFGS:

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

Начало оптимизации на выпуклых множествах. Проекционный GD

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

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

GD с проекцией:

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

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:

- ℓ_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

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

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:

- ℓ_1-шар радиуса R с центром в 0:
  $y^* = -R\mathrm{sign}(x_i)\mathbf{e_i}$, $i = \arg\max_j |x_j|$,
- Симплекс $\triangle = \left\{y \in \mathbb{R}^d \mid y_i \ge 0, \sum_{i=1}^d y_i = R\right\}$:

$$
y^* = R\mathbf{e_i}, \text{ где } i = \arg\min x_i.
$$

Зеркальный спуск

Ставится задача минимизации на множестве $\min_{x \in X} f(x)$.
Идея: обобщить понятия расстояний и проекции, использовав геометрию задачи и выиграв в константе сложности

Дивергенцией Брэгмана (аналог метрики)

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

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

Примеры:

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

$$d(x) = \sum_{i=1}^{d} x_i \log x_i \Longrightarrow V(x, y) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i}.$$

Зеркальный спуск 2

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

Эквивалентная запись

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

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

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

На единичном симплексе:

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

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

где f выпуклая и M-Липшицева.
Субградиентный метод:

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

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 $\gamma = \frac{\|x^0-x^*\|_2}{M\sqrt{K}}$.

Проксимальный оператор

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

Примеры:

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

Свойства:

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

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:

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

Альтернативая запись:

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

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 $\text{prox}_{\gamma r}(\cdot - \gamma \nabla f(\cdot))$

Начало оптимизации с ограничениями. Штрафная функция

Рассмотрим следующую задачу с ограничениями:

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

Аугментация:

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

Пусть все функции являются непрерывными 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

$$X_e^* = \{x \in \mathbb{R}^d \mid \exists x^* \in X^* \ : \ \|x - x^*\|_2 \le e\}.$$

Метод штрафных функций

Алгоритм:

1. решить задачу для текущего ρ,
2. увеличить ρ,
3. использовать предыдущее решение как начальную точку.

Особенности:

- Условная задача превращена в безусловную.
- 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.

ADMM

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

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

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

Алгоритм:

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

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

ADMM: Детали

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

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$
$$\text{s.t.} \quad g_i(x) \leq 0, \quad i = 1, \ldots m.$$

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: $F(x) = -\sum_{i=1}^m \frac{1}{g_i(x)}$;
- Логарифмический барьер: $F(x) = -\sum_{i=1}^m \ln(-g_i(x))$.

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

Метод внутренней точки

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

$$X^*_e = \{x \in G \mid \exists x^* \in X^* \ : \ \|x - x^*\|_2 \leq e\},$$

где X^* – множество решений исходной задачи оптимизации с ограничениями вида неравенств.

Алгоритм:

1. Увеличить $\rho_k > \rho_{k-1}$
2. С помощью некоторого метода решить численно задачу безусловной оптимизации с целевой функцией F_{ρ_k} и стартовой точкой x_k. Гарантировать, что выход метода x_{k+1} будет близок к реальному решению $x^*(\rho_k)$.

Всегда соблюдаем ограничения неравенства !

Седловая задача и Экстреградиент

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

где L непрерывно дифференцируема по обеим группам переменных, выпукла-вогнута: выпукла по x (для любого фиксированного λ) и вогнута по λ (для любого фиксированного x)
Обобщения GD расходятся или сходятся неоптимально ($L(x, y) = xy$).

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

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: $\max_\lambda L(x^k, \lambda) - \min_x L(x, \lambda^k)$

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

Стохастическая оптимизация

Онлайн-постановка:

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

Оффлайн-постановка:

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

SGD:

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

Предположения:

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

SGD Сходимость

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

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

Постоянный шаг:

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

Уменьшающийся шаг: $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$. Плюс: точнее сходимость, минус: потеря линейной сходимости в начале.

Batching:

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

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

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \le \left(1 - \frac{\mu}{L}\right)^k \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}$$

SAGA

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

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

Алгоритм SAGA:

Сгенерировать независимо i_k

$$g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n}\sum_{j=1}^{n} y_j^k$$

$$y_i^{k+1} = \begin{cases} \nabla f_i(x^k), \text{ если } i = i_k \\ y_i^k, \text{ иначе} \end{cases}$$

$$x^{k+1} = x^k - \gamma g^k$$

$\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k$ – «запаздывающая» версия $\nabla f(x^k)$, $\mathbb{E}\left[g^k | x^k\right] = \nabla f(x^k)$.

При $x^k \to x^*$ имеем, что $y_j^k \to \nabla f_j(x^*)$, и $\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k \to \nabla f(x^*) = 0$, $\nabla f_{i_k}(x^k) \to \nabla f_j(x^*)$, 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 итерационная сложность:

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