Универсальные методы для стохастических вариационных неравенств

Климза Антон
Руководитель: А. В. Гасников

Аннотация

В данной статье рассматривается задача оптимизации стохастических вариационных неравенств. Мы предлагаем стохастический вариант универсального проксимального зеркального метода для решения задачи оптимизации. Получены оценки необходимого числа итераций для достижения заданного качества решения вариационного неравенства. Также, мы сравниваем полученный алгоритм с другими популярными оптимизаторами на задаче классификации по картинкам.

1 Введение

Вариационные неравенства нередко возникают в самых разных проблемах оптимизации и имеют многочисленные приложения [1] в математической экономике, теории игр и машинном обучении для задач негладких оптимизаций [2], генеративно-состязательных сетей [3] и обучения с подкреплением [4, 5]. Наиболее известным аналогом градиентного метода для вариационных неравенств является экстраградиентный метод Г.М. Корпелевич [6]. Одним из современных вариантов экстраградиентного метода является проксимальный зеркальный метод А.С. Немировского [7].

Задачу стохастической выпуклой оптимизации уже разбирали в статье [9], в которой предлагается универсальный метод для решения монотонных стохастических вариационных неравенств на базе проксимального зеркального метода. По сути, используется стандартный проксимальный зеркальный метод, в котором L предлагается выбирать специальным образом, схожим со способом, использующимся в Adagrad. Однако этот метод не является полностью адаптивным, поскольку, так же как и в Adagrad, в стратегии выбора шага существенно используется информация о размере решения. Полностью адаптивный метод решения гладких стохастических монотонных вариационных неравенств был построен (с небольшими оговорками) в работе [10] и в работе [12] для неточного аракула.

В новой статье [8] авторы предлагают свой универсальный градиентный спуск для задач стохастической выпуклой оптимизации. Мы предлагаем применение этого метода для стохастических вариационных неравенств, в частности для седловых задач. Такие постановки, например, возникают в задачах состязательного обучения. Преимущества универсального градиентного спуска в том, что он сам настраивается на гладкость задачи и не требует параметров на входе.

2 Постановка задачи

Для некоторого оператора $g : Q \to \mathbb{R}^n$, заданного на выпуклом компакте $Q \in \mathbb{R}^n$, будем рассматривать сильные вариационные неравенства вида:

$$\langle g(x^*), x^* - x \rangle \leq 0.$$

Отметим, что в этом неравенстве требуется найти решение вариационного неравенства $x^* \in Q$, для которого

$$\max_{x \in Q} \langle g(x^*), x^* - x \rangle \leq 0.$$

В случае монотонного поля наш подход позволяет рассматривать также слабые вариационные неравенства

$$\langle g(x), x^* - x \rangle \leq 0,$$

в котором требуется найти $x^* \in Q$, такое, что неравенство верно при всех $x \in Q$. Обозначим:

$$Gap(x^*) = \max_{x \in Q} \langle g(x), x^* - x \rangle ,$$

и будем считать x^* - ε-решением вариационного неравенства, если $Gap(x^*) \leq \varepsilon$.
Также предполагаем, что вариационное неравенство монотонное (при $g = \nabla f$ равносильно Q - выпуклое):

$$\langle g(x) - g(y), x - y \rangle \geq 0 \quad \forall x, y \in Q,$$

Q удовлетворяет условию Гёльдера:

$$\exists \nu \in [0, 1], L_\nu \geq 0 : \; \|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu \quad \forall x, y \in Q,$$

и Q ограничено константой $D \geq \max\limits_{x,y \in Q} \|x - y\|$.
В стохастическом случаи предполагаем:

$$E_\xi g(x, \xi) = g(x),$$

$$\exists \sigma : \; E_\xi \|g(x, \xi) - g(x)\|^2 \leq \sigma^2.$$

3 Универсальный проксимальный зеркальный метод

В проксимальном зеркальном методе мы рассматриваем шаг:

$$w_k = arg \min_{x \in Q} \left(\langle g(z_k), x - z_k \rangle + L_k \frac{1}{2} \|z_k - x\|^2 \right),$$

$$z_{k+1} = arg \min_{x \in Q} \left(\langle g(w_k), x - w_k \rangle + L_k \frac{1}{2} \|z_k - x\|^2 \right).$$

В обычном проксимальном зеркальном методе мы подсчитываем константу L_k для $k = 0$ и не изменяем. В универсальном методе мы хотим пересчитывать L_k по простой формуле, тем самым получая более точное приближение константы Гёльдера в окрестности текущей точки. Сначала докажем следующую лемму:

Лемма 1. *Оператор* $h_k(z^*) = \langle g(w), z^* - w\rangle + L_k\frac{1}{2}\|z - z^*\|^2$ *- сильно выпуклый с константой* L_k*:*

Доказательство.

$$h_k(\alpha x + (1-\alpha)y) = \langle g(w), \alpha x + (1-\alpha)y - w\rangle + L_k\frac{1}{2}\|\alpha x + (1-\alpha)y - z\|^2 =$$

$$= \alpha \langle g(w), x - w\rangle + (1-\alpha)\langle g(w), y - w\rangle + L_k\frac{1}{2}\|\alpha x + (1-\alpha)y - z\|^2 =$$

$$= \alpha h_k(x) + (1-\alpha)h_k(y) + L_k\frac{1}{2}\|\alpha x + (1-\alpha)y - z\|^2 - \alpha L_k\frac{1}{2}\|x - z\|^2 - (1-\alpha)L_k\frac{1}{2}\|y - z\|^2 =$$

$$= \alpha h_k(x) + (1-\alpha)h_k(y) + L_k\frac{1}{2}\|x - y\|^2 +$$

$$L_k\frac{\alpha^2 - \alpha - 1}{2}\|x\|^2 + L_k\frac{(1-\alpha)^2 - (1-\alpha) - 1}{2}\|y\|^2 + L_k(\alpha(1-\alpha) - 1)\langle x, y\rangle =$$

$$= \alpha h_k(x) + (1-\alpha)h_k(y) + L_k\frac{1}{2}\|x - y\|^2 + L_k\frac{\alpha^2 - \alpha - 1}{2}f(x,y),$$

где $f(x,y) = \|x\|^2 + \|y\|^2 - 2\langle x, y\rangle \geq 0 \;\forall x, y$. Тогда $h_k(\alpha x + (1-\alpha)y) \geq \alpha h_k(x) + (1-\alpha)h_k(y) + L_k\frac{1}{2}\|x - y\|^2$, а значит оператор $h_k(z^*)$ - сильно выпуклый с константой L_k. □

По лемме 1 получаем, что для $x = arg\min\limits_{z^* \in Q} h_k(z^*)$ выполнено:

$$h_k(z^*) \geq h_k(x) + L_k\frac{1}{2}\|z^* - x\|^2,$$

Тогда получим неравенство:

$$\langle g(w_k), z^* - w_k\rangle + L_k\frac{1}{2}\|z_k - z^*\|^2 \geq \langle g(w_k), z_{k+1} - w_k\rangle + L_k\frac{1}{2}\|z_k - z_{k+1}\|^2 + L_k\frac{1}{2}\|z_{k+1} - z^*\|^2.$$

$$\langle g(w_k), w_k - z^*\rangle + L_k\frac{1}{2}\|z_{k+1} - z^*\|^2 \leq \langle g(w_k), w_k - z_{k+1}\rangle - L_k\frac{1}{2}\|z_k - z_{k+1}\|^2 + L_k\frac{1}{2}\|z_k - z^*\|^2.$$

Мы получили оценку для L_k, теперь, чтобы получить формулу пересчёта L_{k+1}, мы добавим его в неравенство. К обоим частям прибавим $L_{k+1}\frac{1}{2}\|z_{k+1} - z^*\|^2$:

$$\langle g(w_k), w_k - z^*\rangle + L_{k+1}\frac{1}{2}\|z_{k+1} - z^*\|^2 \leq$$

$$\leq \langle g(w_k), w_k - z_{k+1}\rangle - L_k\frac{1}{2}\|z_k - z_{k+1}\|^2 + (L_{k+1} - L_k)\frac{1}{2}\|z_{k+1} - z^*\|^2 + L_k\frac{1}{2}\|z_k - z^*\|^2.$$

Тогда при $L_{k+1} \geq L_k$:

$$\langle g(w_k), w_k - z^*\rangle + L_{k+1}\frac{1}{2}\|z_{k+1} - z^*\|^2 \leq$$

$$\leq \langle g(w_k), w_k - z_{k+1}\rangle - L_k\frac{1}{2}\|z_k - z_{k+1}\|^2 + (L_{k+1} - L_k)\frac{D^2}{2} + L_k\frac{1}{2}\|z_k - z^*\|^2.$$

Будем искать L_{k+1} такое, что:

$$(L_{k+1} - L_k)\frac{D^2}{2} = \left|\langle g(w_k), w_k - z_{k+1}\rangle - L_k\frac{1}{2}\|z_k - z_{k+1}\|^2\right|_+.$$

Тогда:

$$\langle g(w_k), w_k - z^* \rangle + L_{k+1}\frac{1}{2}\|z_{k+1} - z^*\|^2 \leq D^2(L_{k+1} - L_k) + L_k\frac{1}{2}\|z_k - z^*\|^2.$$

Через телескопическую сумму получаем:

$$\sum_{i=0}^{k} \langle g(w_k), w_k - z^* \rangle + L_{k+1}\frac{1}{2}\|z_{k+1} - z^*\|^2 \leq D^2(L_{k+1} - L_0) + L_0\frac{1}{2}\|z_0 - z^*\|^2,$$

$$\frac{1}{k}\sum_{i=0}^{k} \langle g(w_k), w_k - z^* \rangle \leq \frac{D^2 L_{k+1}}{k}.$$

Тогда для монотонного вариационного неравенства получаем $\langle g(z^*) - g(w_k), w_k - z^* \rangle \leq 0$, поэтому для $w = \frac{1}{k}\sum\limits_{i=0}^{k} w_k$:

$$Gap(w) = \max_{z^* \in Q} \langle g(z^*), w - z^* \rangle = \max_{z^* \in Q} \frac{1}{k}\sum_{i=0}^{k} \left(\langle g(w_k), w_k - z^* \rangle + \langle g(z^*) - g(w_k), w_k - z^* \rangle \right),$$

$$Gap(w) \leq \frac{D^2 L_{k+1}}{k}.$$

Мы получили оценку сходимости метода, при данном выборе L_{k+1}. Теперь необходимо доказать, что метод сходится, для чего хотим $L_{k+1} = o(k)$. Для этого возьмём немного другое L_{k+1} по формуле:

$$(L_{k+1} - L_k)\frac{D^2}{2} = \left| \langle g(w_k), w_k - z_{k+1} \rangle - L_{k+1}\frac{1}{2}\|z_k - z_{k+1}\|^2 \right|_+.$$

Для такого пересчёта будет выполнятся другое неравенство и сходимость метода немного ухудшится, однако с такой заменой нам удастся оценить L_{k+1}:

$$\langle g(w_k), w_k - z^* \rangle + L_{k+1}\frac{1}{2}\|z_{k+1} - z^*\|^2 \leq$$

$$\leq (D^2 + \|z_{k+1} - z_k\|^2)(L_{k+1} - L_k) + L_k\frac{1}{2}\|z_k - z^*\|^2 \leq 2D^2(L_{k+1} - L_k) + L_k\frac{1}{2}\|z_k - z^*\|^2.$$

Аналогично получим:

$$Gap(w) \leq \frac{2D^2 L_{k+1}}{k}.$$

Лемма 2.

$$L_{k+1} \leq \left(\frac{8k}{D^2}\right)^{\frac{1-\nu}{2}} L_\nu.$$

Доказательство. Если правая часть отрицательная, то получаем $L_{k+1} = L_k$, иначе:

$$(L_{k+1} - L_k)\frac{D^2}{2} = \langle g(w_k), w_k - z_{k+1} \rangle - L_{k+1}\frac{1}{2}\|z_k - z_{k+1}\|^2.$$

Воспользуемся Гёльдеровостью оператора g:

$$\langle g(w_k) - g(z_k), w_k - z_{k+1}\rangle \le L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\|,$$

$$\langle g(w_k), w_k - z_{k+1}\rangle \le \langle g(z_k), w_k - z_{k+1}\rangle + L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\|.$$

Снова воспользуемся неравенством $h_k(z^*) \ge h_k(x) + L_k \frac{1}{2}\|z^* - x\|^2$ для оператора h, где в качестве z^* возьмём z_{k+1}:

$$\langle g(z_k), w_k - z_k\rangle + L_k \frac{1}{2}\|z_k - w_k\|^2 + L_k \frac{1}{2}\|z_{k+1} - w_k\|^2 \le \langle g(z_k), z_{k+1} - z_k\rangle + L_k \frac{1}{2}\|z_k - z_{k+1}\|^2.$$

Применим его для оценки $\langle g(z_k), w_k - z_{k+1}\rangle$:

$$\langle g(z_k), w_k - z_{k+1}\rangle \le L_k \frac{1}{2}\|z_k - z_{k+1}\|^2 - L_k \frac{1}{2}\|z_k - w_k\|^2 - L_k \frac{1}{2}\|z_{k+1} - w_k\|^2,$$

$$\langle g(w_k), w_k - z_{k+1}\rangle \le L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| + L_k \frac{1}{2}\|z_k - z_{k+1}\|^2 - L_k \frac{1}{2}\|z_k - w_k\|^2 - L_k \frac{1}{2}\|z_{k+1} - w_k\|^2.$$

Применяем неравенство для оценки L_{k+1}:

$$(L_{k+1} - L_k)\frac{1}{2}D^2 \le L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| + (L_k - L_{k+1})\frac{1}{2}\|z_k - z_{k+1}\|^2 - L_k \frac{1}{2}\|z_k - w_k\|^2 - L_k \frac{1}{2}\|z_{k+1} - w_k\|^2,$$

$$(L_{k+1} - L_k)\frac{1}{2}(D^2 + \|z_k - z_{k+1}\|^2 - \|z_k - w_k\|^2 - \|z_{k+1} - w_k\|^2) \le$$

$$\le L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| - L_{k+1}\frac{1}{2}\left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2\right).$$

Обозначим $R = D^2 + \|z_k - z_{k+1}\|^2 - \|z_k - w_k\|^2 - \|z_{k+1} - w_k\|^2$, тогда:

$$(L_{k+1} - L_k)\frac{R}{2} \le L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| - L_{k+1}\frac{1}{2}\left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2\right).$$

Найдём максимум функции $f(x, y) = L_\nu x^\nu y - \frac{1}{2}L_{k+1}(x^2 + y^2)$ при $\nu \in (0, 1)$:

$$\frac{\partial f}{\partial x} = \nu L_\nu x^{\nu-1} y - L_{k+1}x,$$

$$\frac{\partial f}{\partial y} = L_\nu x^\nu - L_{k+1}y.$$

Заметим, что в точке $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$ фукция будет принимать максимальное значение. Тогда:

$$\nu L_\nu x^{\nu-1} y - L_{k+1}x = 0, \;\Rightarrow\; y = \frac{L_{k+1}x^{2-\nu}}{\nu L_\nu}.$$

$$L_\nu x^\nu - L_{k+1}y = 0, \;\Rightarrow\; y = \frac{L_\nu x^\nu}{L_{k+1}}.$$

$$\frac{L_{k+1}x^{2-\nu}}{\nu L_\nu} = \frac{L_\nu x^\nu}{L_{k+1}}, \;\Rightarrow\; x = \left(\frac{\sqrt{\nu}L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}}, \; y = \left(\frac{\nu^{\frac{\nu}{2}}L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}}.$$

Получаем:

$$f(x,y) \le L_\nu \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}}\right)^{\frac{\nu}{1-\nu}} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}} - \frac{1}{2} L_{k+1} \left(\left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}}\right)^{\frac{2}{1-\nu}} + \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{2}{1-\nu}}\right) =$$

$$= \frac{\nu^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}}}{L_{k+1}^{\frac{1+\nu}{1-\nu}}} \left(1 - \frac{\nu^{\frac{1}{\nu}}}{2} - \frac{1}{2}\right) = \frac{\nu^{\frac{\nu}{1-\nu}} (1 - \nu^{\frac{1}{\nu}}) L_\nu^{\frac{2}{1-\nu}}}{2 L_{k+1}^{\frac{1+\nu}{1-\nu}}}.$$

Пусть $p = \frac{1+\nu}{1-\nu}$, тогда:

$$(L_{k+1} - L_k) \frac{R}{2} \le \frac{\nu^{\frac{\nu}{1-\nu}} (1 - \nu^{\frac{1}{\nu}}) L_\nu^{p+1}}{2 L_{k+1}^p},$$

$$(p+1) L_{k+1}^p (L_{k+1} - L_k) \le \frac{\nu^{\frac{\nu}{1-\nu}} (1 - \nu^{\frac{1}{\nu}})(p+1) L_\nu^{p+1}}{R} = \alpha.$$

Заметим, что:

$$(p+1) L_{k+1}^p (L_{k+1} - L_k) \ge (p+1) \int\limits_{L_k}^{L_{k+1}} t^p dt = L_{k+1}^{p+1} - L_k^{p+1}.$$

Тогда:

$$L_{k+1}^{p+1} - L_k^{p+1} \le \alpha,$$

и по телескопической сумме получаем:

$$L_{k+1}^{p+1} \le k\alpha + L_0^{p+1}.$$

Для достаточно большого k выполнено $k\alpha \ge L_0^{p+1}$, тогда:

$$L_{k+1} \le (2k\alpha)^{\frac{1}{p+1}} = \left(2k \frac{\nu^{\frac{\nu}{1-\nu}} (1 - \nu^{\frac{1}{\nu}})(p+1) L_\nu^{p+1}}{R}\right)^{\frac{1}{p+1}} = k^{\frac{1-\nu}{2}} \frac{\nu^{\frac{\nu}{2}} (1 - \nu^{\frac{1}{\nu}})^{\frac{1-\nu}{2}} L_\nu}{\left(\frac{1-\nu}{4}\right)^{\frac{1-\nu}{2}} R^{\frac{1-\nu}{2}}} \le \left(\frac{4k}{R}\right)^{\frac{1-\nu}{2}} L_\nu.$$

По тождеству параллелограмма:

$$R = D^2 + \|z_k - z_{k+1}\|^2 - \|z_k - w_k\|^2 - \|z_{k+1} - w_k\|^2 = D^2 + \frac{1}{2}\|z_k - z_{k+1}\|^2 - \frac{1}{2}\|z_k + z_{k+1} - 2w_k\|^2 \ge \frac{D^2}{2}.$$

Тогда:

$$L_{k+1} \le \left(\frac{8k}{D^2}\right)^{\frac{1-\nu}{2}} L_\nu.$$

□

Получаем оценку скорости сходимости метода:

$$Gap\left(\frac{1}{k}\sum_{i=0}^{k} w_k\right) \le \frac{16 L_\nu D^{1+\nu}}{(8k)^{\frac{1+\nu}{2}}} = \varepsilon,$$

$$k = O\left(\inf_{\nu \in [0,1]} \left(\frac{16 L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} \frac{D^2}{8}\right).$$

Algorithm 1 Универсальный проксимальный зеркальный метод (UMP)

Set $z_0 = arg\min_{u\in Q} d(u),\ L_0 = \|g(z_0)\|$.
for $k = 0, 1, \dots$ **do**
  $w_k = arg\min_{x\in Q}\left(\langle g(z_k), x\rangle + L_k\frac{1}{2}\|z_k - x\|^2\right)$.
  $z_{k+1} = arg\min_{x\in Q}\left(\langle g(w_k), x\rangle + L_k\frac{1}{2}\|z_k - x\|^2\right)$.
  $L_{k+1} = L_k + \max\left(0, \frac{2\langle g(w_k), w_k - z_{k+1}\rangle - L_k\|w_k - z_{k+1}\|^2}{D^2 + \|w_k - z_{k+1}\|^2}\right)$.
end for

Теперь рассмотрим стохастический случай. У нас $g(x) = E_\xi g(x,\xi)$, но пересчёт L_{k+1}, w_k и z_{k+1} мы можем получить только используя шумный оператор $g(x,\xi)$. Поэтому:

$$(L_{k+1} - L_k)\frac{D^2}{2} = \left|\langle g(w_k,\xi_k), w_k - z_{k+1}\rangle - L_{k+1}\frac{1}{2}\|z_k - z_{k+1}\|^2\right|_+ .$$

Повторим все оценки для оператора $g(x,\xi)$ вместо $g(x)$ и получим:

$$\langle g(w_k,\xi_k), w_k - z^*\rangle + L_{k+1}\frac{1}{2}\|z_{k+1} - z^*\|^2 \le 2D^2(L_{k+1} - L_k) + L_k\frac{1}{2}\|z_k - z^*\|^2.$$

Так как ξ_k независимо с w_k и z^*, то $E_{\xi_k}\langle g(w_k,\xi_k), w_k - z^*\rangle = \langle g(w_k), w_k - z^*\rangle$. Тогда получаем:

$$E_\xi(Gap(w)) \le \frac{2D^2 E_\xi(L_{k+1})}{k}.$$

Докажем следующую лемму:

Лемма 3.

$$E_\xi(L_{k+1}) \le 2\left(\frac{8k}{D^2}\right)^{\frac{1-\nu}{2}} L_\nu + \frac{2\sqrt{2}}{D}\sigma\sqrt{k} + L_0.$$

Доказательство. Если правая часть отрицательная, то получаем $L_{k+1} = L_k$, иначе:

$$(L_{k+1} - L_k)\frac{D^2}{2} = \langle g(w_k,\xi_k), w_k - z_{k+1}\rangle - L_{k+1}\frac{1}{2}\|z_k - z_{k+1}\|^2 =$$

$$= \langle g(w_k,\xi_k) - g(w_k), w_k - z_{k+1}\rangle + \langle g(w_k), w_k - z_{k+1}\rangle - L_{k+1}\frac{1}{2}\|z_k - z_{k+1}\|^2.$$

Аналогично лемме 2 получаем:

$$(L_{k+1} - L_k)\frac{R}{2} \le$$

$$\le L_\nu\|w_k - z_k\|^\nu\|w_k - z_{k+1}\| - L_{k+1}\frac{1}{2}\left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2\right) + \langle g(w_k,\xi_k) - g(w_k), w_k - z_{k+1}\rangle.$$

Обозначим $\delta_k = \|g(w_k,\xi_k) - g(w_k)\|$, тогда:

$$(L_{k+1} - L_k)\frac{R}{2} \le L_\nu\|w_k - z_k\|^\nu\|w_k - z_{k+1}\| - L_{k+1}\frac{1}{2}\left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2\right) + \delta_k\|w_k - z_{k+1}\|.$$

Найдём максимум функции $f(x,y) = L_\nu x^\nu y + \delta_k y - \frac{1}{2}L_{k+1}(x^2 + y^2)$ при $\nu \in (0,1)$:

$$\frac{\partial f}{\partial x} = \nu L_\nu x^{\nu-1}y - L_{k+1}x,$$

$$\frac{\partial f}{\partial y} = L_\nu x^\nu - L_{k+1} y + \delta_k.$$

Заметим, что в точке $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$ фукция будет принимать максимальное значение. Тогда:

$$\nu L_\nu x^{\nu-1} y - L_{k+1} x = 0, \ \Rightarrow \ y = \frac{L_{k+1} x^{2-\nu}}{\nu L_\nu}.$$

$$L_\nu x^\nu - L_{k+1} y + \delta_k = 0, \ \Rightarrow \ y = \frac{L_\nu x^\nu + \delta_k}{L_{k+1}}.$$

$$\frac{L_{k+1} x^{2-\nu}}{\nu L_\nu} = \frac{L_\nu x^\nu}{L_{k+1}} + \frac{\delta_k}{L_{k+1}}.$$

Подставим $x = C_k \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}}$, тогда $y = C_k^{2-\nu} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}}$ и выполнено равенство:

$$C_k^{2-\nu} - C_k^\nu - \frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\frac{\nu}{2}} L_\nu}\right)^{\frac{1}{1-\nu}} = 0.$$

Так как $\delta_k \geq 0$, то максимум функции $f(x,y)$ достигается при $y \geq 0$, тогда $C_k \geq 0$. Теперь подставим $x = C_k \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}}$ и $y = C_k^{2-\nu} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}}$ в $f(x,y)$:

$$f(x,y) = C_k^2 L_\nu \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}}\right)^{\frac{\nu}{1-\nu}} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}} - \frac{1}{2} L_{k+1} \left(C_k^2 \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}}\right)^{\frac{2}{1-\nu}} + C_k^{4-2\nu} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{2}{1-\nu}} \right) +$$

$$+ \delta_k C_k^{2-\nu} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}} = \frac{\nu^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}}}{L_{k+1}^{\frac{1+\nu}{1-\nu}}} C_k^2 \left(1 - \frac{\nu^{\frac{1}{\nu}}}{2} - \frac{C_k^{2-2\nu}}{2}\right) + \delta_k C_k^{2-\nu} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}} =$$

$$= \frac{\nu^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}}}{2 L_{k+1}^{\frac{1+\nu}{1-\nu}}} C_k^2 \left(1 - \nu^{\frac{1}{\nu}} - \frac{\frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\frac{\nu}{2}} L_\nu}\right)^{\frac{1}{1-\nu}}}{C_k^\nu}\right) + \delta_k C_k^{2-\nu} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}} =$$

$$= \frac{\nu^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}} (1 - \nu^{\frac{1}{\nu}})}{2 L_{k+1}^{\frac{1+\nu}{1-\nu}}} C_k^2 + \frac{\delta_k C_k^{2-\nu}}{2} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}}.$$

Рассмотрим 2 случая:

1) Пусть $\frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\frac{\nu}{2}} L_\nu}\right)^{\frac{1}{1-\nu}} \leq C_k^\nu$, тогда $0 = C_k^{2-\nu} - C_k^\nu - \frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\frac{\nu}{2}} L_\nu}\right)^{\frac{1}{1-\nu}} \geq C_k^{2-\nu} - 2C_k^\nu$, поэтому $C_k \leq 2^{\frac{1}{2-2\nu}}$, значит:

$$f(x,y) = \frac{\nu^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}} (1 - \nu^{\frac{1}{\nu}})}{2 L_{k+1}^{\frac{1+\nu}{1-\nu}}} C_k^2 + \frac{\delta_k C_k^{2-\nu}}{2} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}} \leq$$

$$\leq \frac{(2\nu)^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}} (1 - \nu^{\frac{1}{\nu}})}{L_{k+1}^{\frac{1+\nu}{1-\nu}}} + 2^{\frac{\nu}{1-\nu}} L_{k+1} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{2}{1-\nu}} = \frac{(2\nu)^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}}}{L_{k+1}^{\frac{1+\nu}{1-\nu}}} \left(2 - \nu^{\frac{1}{\nu}}\right)$$

Пусть $p = \frac{1+\nu}{1-\nu}$, тогда аналогично лемме 2 выполнено $L_{k+1}^{p+1} \le \alpha_k + L_k^{p+1}$, где:

$$\alpha_k = \frac{2(p+1)(2\nu)^{\frac{\nu}{1-\nu}} L_\nu^{p+1}}{R} \left(2 - \nu^{\frac{1}{\nu}}\right).$$

Рассмотрим второй случай:
2) $\frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\frac{\nu}{2}} L_\nu}\right)^{\frac{1}{1-\nu}} \ge C_k^\nu$, тогда $0 = C_k^{2-\nu} - C_k^\nu - \frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\frac{\nu}{2}} L_\nu}\right)^{\frac{1}{1-\nu}} \ge C_k^{2-\nu} - \frac{2\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\frac{\nu}{2}} L_\nu}\right)^{\frac{1}{1-\nu}}$,
поэтому $C_k \le \left(\frac{2\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\frac{\nu}{2}} L_\nu}\right)^{\frac{1}{1-\nu}}\right)^{\frac{1}{2-\nu}}$, значит:

$$f(x,y) = \frac{\nu^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}} (1 - \nu^{\frac{1}{\nu}})}{2 L_{k+1}^{\frac{1+\nu}{1-\nu}}} C_k^2 + \frac{\delta_k C_k^{2-\nu}}{2} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}} \le$$

$$\le \frac{\delta_k C_k^{2-\nu} (2 - \nu^{\frac{1}{\nu}})}{2} \left(\frac{\nu^{\frac{\nu}{2}} L_\nu}{L_{k+1}}\right)^{\frac{1}{1-\nu}} \le \frac{\delta_k^2 (2 - \nu^{\frac{1}{\nu}})}{L_{k+1}}.$$

Пусть $q = 1$, тогда аналогично лемме 2 выполнено $L_{k+1}^{q+1} \le \beta_k + L_k^{q+1}$, где:

$$\beta_k = \frac{2\delta_k^2 (2 - \nu^{\frac{1}{\nu}})}{R}.$$

Докажем следующую лемму:

Лемма 4. *Даны последовательности положительных чисел x_k, α_k, β_k. Пусть для любого k существуют $p, q \ge 1$ такие, что выполнено хотя бы одно из двух условий:*

$$\left[\begin{array}{l} x_{k+1}^p \le \alpha_k + x_k^p, \\ x_{k+1}^q \le \beta_k + x_k^q. \end{array}\right.$$

Тогда $x_{k+1} \le \left(\sum\limits_{i=0}^{k} \alpha_i\right)^{\frac{1}{p}} + \left(\sum\limits_{i=0}^{k} \beta_i\right)^{\frac{1}{q}} + x_0$.

Доказательство. Докажем по индукции:
База:

- Пусть $x_1^p \le \alpha_0 + x_0^p$, тогда $x_1 \le (\alpha_0 + x_0^p)^{\frac{1}{p}} \le (\alpha_0^{\frac{1}{p}} + x_0) + \beta_0^{\frac{1}{q}}$.

- Пусть $x_1^q \le \beta_0 + x_0^q$, тогда $x_1 \le (\beta_0 + x_0^q)^{\frac{1}{q}} \le \alpha_0^{\frac{1}{p}} + (\beta_0^{\frac{1}{q}} + x_0)$.

Переход:

Пусть известно, что $x_k \le \left(\sum\limits_{i=0}^{k-1} \alpha_i\right)^{\frac{1}{p}} + \left(\sum\limits_{i=0}^{k-1} \beta_i\right)^{\frac{1}{q}} + x_0$, тогда:
Если $x_{k+1}^p \le \alpha_k + x_k^p$, то:

$$x_{k+1} \le \left(\alpha_k + \left(\left(\sum_{i=0}^{k-1} \alpha_i\right)^{\frac{1}{p}} + \left(\sum_{i=0}^{k-1} \beta_i\right)^{\frac{1}{q}} + x_0\right)^p\right)^{\frac{1}{p}}.$$

Рассмотрим вектора $(\alpha_k, 0, 0)$ и $(\sum\limits_{i=0}^{k-1} \alpha_i, \sum\limits_{i=0}^{k-1} \beta_i, x_0^p)$. Тогда по обратному неравенству треугольника для пространства Минковского:

$$\left(\alpha_k^{\frac{1}{p}} + 0^{\frac{1}{p}} + 0^{\frac{1}{p}}\right)^p + \left(\left(\sum_{i=0}^{k-1} \alpha_i\right)^{\frac{1}{p}} + \left(\sum_{i=0}^{k-1} \beta_i\right)^{\frac{1}{p}} + (x_0^p)^{\frac{1}{p}}\right)^p \leq$$

$$\leq \left(\left(\alpha_k + \sum_{i=0}^{k-1} \alpha_i\right)^{\frac{1}{p}} + \left(0 + \sum_{i=0}^{k-1} \beta_i\right)^{\frac{1}{p}} + (0 + x_0^p)^{\frac{1}{p}}\right)^p.$$

Тогда получаем:

$$x_{k+1} \leq \left(\sum_{i=0}^{k} \alpha_i\right)^{\frac{1}{p}} + \left(0 + \sum_{i=0}^{k-1} \beta_i\right)^{\frac{1}{p}} + x_0 \leq \left(\sum_{i=0}^{k} \alpha_i\right)^{\frac{1}{p}} + \left(\sum_{i=0}^{k} \beta_i\right)^{\frac{1}{p}} + x_0.$$

Если $x_{k+1}^q \leq \beta_k + x_k^q$, то:

$$x_{k+1} \leq \left(\beta_k + \left(\left(\sum_{i=0}^{k-1} \alpha_i\right)^{\frac{1}{p}} + \left(\sum_{i=0}^{k-1} \beta_i\right)^{\frac{1}{q}} + x_0\right)^q\right)^{\frac{1}{q}}.$$

Рассмотрим вектора $(0, \beta_k, 0)$ и $(\sum\limits_{i=0}^{k-1} \alpha_i, \sum\limits_{i=0}^{k-1} \beta_i, x_0^p)$. Тогда по обратному неравенству треугольника для пространства Минковского:

$$\left(0^{\frac{1}{p}} + \beta_k^{\frac{1}{p}} + 0^{\frac{1}{p}}\right)^p + \left(\left(\sum_{i=0}^{k-1} \alpha_i\right)^{\frac{1}{p}} + \left(\sum_{i=0}^{k-1} \beta_i\right)^{\frac{1}{p}} + (x_0^p)^{\frac{1}{p}}\right)^p \leq$$

$$\leq \left(\left(0 + \sum_{i=0}^{k-1} \alpha_i\right)^{\frac{1}{p}} + \left(\beta_k + \sum_{i=0}^{k-1} \beta_i\right)^{\frac{1}{p}} + (0 + x_0^p)^{\frac{1}{p}}\right)^p.$$

Тогда получаем:

$$x_{k+1} \leq \left(0 + \sum_{i=0}^{k-1} \alpha_i\right)^{\frac{1}{p}} + \left(\sum_{i=0}^{k} \beta_i\right)^{\frac{1}{p}} + x_0 \leq \left(\sum_{i=0}^{k} \alpha_i\right)^{\frac{1}{p}} + \left(\sum_{i=0}^{k} \beta_i\right)^{\frac{1}{p}} + x_0.$$

$\square$

Тогда по лемме 4 получаем:

$$L_{k+1} \leq \left(k \frac{2(p+1)(2\nu)^{\frac{\nu}{1-\nu}} L_\nu^{p+1}}{R} \left(2 - \nu^{\frac{1}{\nu}}\right)\right)^{\frac{1}{p+1}} + \sqrt{\frac{2(2-\nu^{\frac{1}{\nu}})}{R} \sum_{i=1}^{k} \delta_i^2} + L_0 \leq$$

$$\leq 2^{\frac{2-\nu}{2}} k^{\frac{1-\nu}{2}} \frac{\nu^{\frac{\nu}{2}} L_\nu}{(1-\nu)^{\frac{1-\nu}{2}} R^{\frac{1-\nu}{2}}} \left(2 - \nu^{\frac{1}{\nu}}\right)^{\frac{1-\nu}{2}} + \frac{2}{\sqrt{R}} \sqrt{\sum_{i=1}^{k} \delta_i^2} + L_0 \leq$$

$$\le 2\left(\frac{4k}{R}\right)^{\frac{1-\nu}{2}} L_\nu + \frac{2}{\sqrt{R}}\sqrt{\sum_{i=1}^{k}\delta_i^2} + L_0.$$

По неравенству Йенсена: $E_\xi\left(\sum\limits_{i=1}^{k}\delta_i^2\right)^{\frac{1}{2}} \le \left(\sum\limits_{i=1}^{k}E_\xi\delta_i^2\right)^{\frac{1}{2}} = \sigma\sqrt{k}$, тогда:

$$E_\xi(L_{k+1}) \le 2\left(\frac{8k}{D^2}\right)^{\frac{1-\nu}{2}} L_\nu + \frac{2\sqrt{2}}{D}\sigma\sqrt{k} + L_0.$$

$\square$

Получаем оценку скорости сходимости метода в стохастическом случаи:

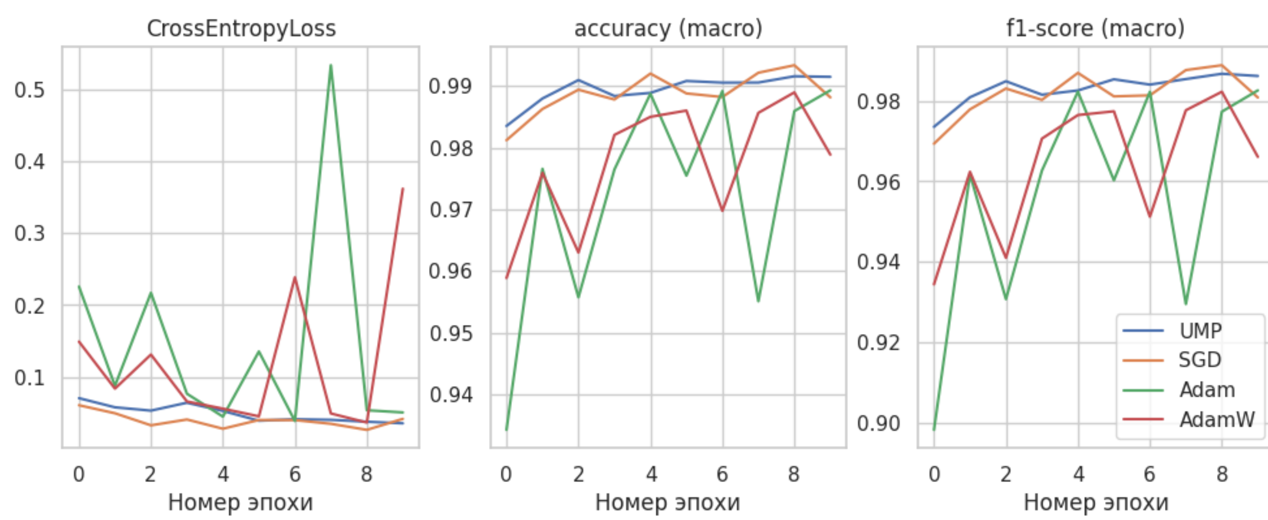
$$E_\xi\left(Gap\left(\frac{1}{k}\sum_{i=0}^{k}w_k\right)\right) \le 32\left(\frac{D^2}{8k}\right)^{\frac{1+\nu}{2}} L_\nu + \frac{4\sqrt{2}D}{\sqrt{k}}\sigma + \frac{2D^2L_0}{k}.$$

4 Вычислительный эксперимент

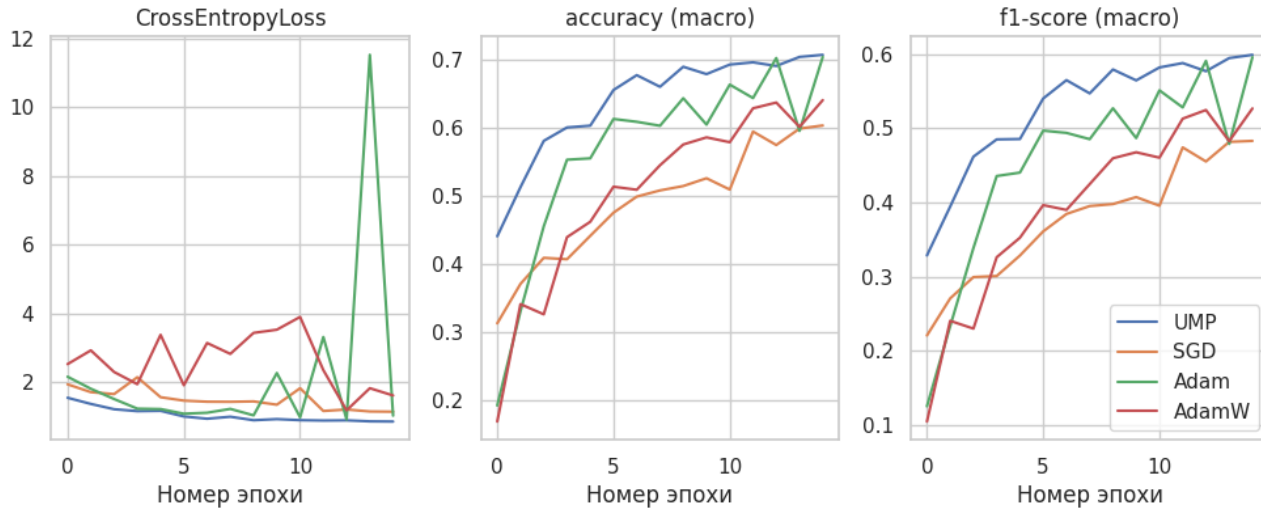
Для сравнения работы универсального проксимального зеркального метода с другими оптимизаторами обучим свёрточную нейеронную сеть resnet50 реализованную в библиотеке pytorch на основе статьи [11] для классификации изображений датасетов MNIST и CIFAR10 также предоставленных библиотекой pytorch.

MNIST - это датасет образцов рукописного написания цифр, содержащий в себе 60000 трейновых и 10000 тестовых чёрно-белых картинок размера 28x28, каждая подписана соответствующей ей цифрой.

CIFAR10 - это датасет образцов объектов из 10 классов (самолёт, автомобиль, птица, кот, олень, собака, лягушка, лошадь, корабль и грузовик), содержащий в себе 50000 трейновых и 10000 тестовых цветных картинок размера 32x32, каждая подписана соответствующим классом.



Сравнение метрик оптимизаторов на валидации при обучении resnet50 датасете MNIST



Сравнение метрик оптимизаторов на валидации при обучении resnet50 датасете CIFAR10

В качестве оптимизаторов рассмотрим SGD, Adam, AdamW и наш универсальный проксимальнй зеркальный метод UMP.

По графикам видим, что на датасете MNIST наш метод получил такое же качество, как и SGD, и на датасете CIFAR10 наш метод показал наилучшие результаты. Также наш метод более устойчив к шуму, что видно по графикам лосса. Данный эксперимент показывает, что наш универсальный проксимальнй зеркальный метод может быть успешно использован для задач классической минимизации.

5 Заключение

В данной работе мы показываем известные улучшения проксимального зеркального метода [13] и предлагаем свой универсальный вариант, основанный на идеях универсального градиентного спуска из статьи [8]. Мы доказываем теоретическую сходимость универсального проксимального зеркального метода и получаем оценку скорости сходимости:

$$O\left(\inf_{\nu\in[0,1]}\left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}}\frac{D^2}{8}\right).$$

В стохастическом случаи мы получаем следующую оценку скорости сходимости метода:

$$E_\xi\left(Gap\left(\frac{1}{k}\sum_{i=0}^{k}w_k\right)\right)\leq 32\left(\frac{D^2}{8k}\right)^{\frac{1+\nu}{2}}L_\nu+\frac{4\sqrt{2}D}{\sqrt{k}}\sigma+\frac{2D^2L_0}{k}.$$

Также мы сравниваем работу метода с известными оптимизаторами на классической задаче минимизации.

В будущем планируется провести больше экспериментов с другими моделями и сравнить работу универсального проксимального зеркального метода с другими методами на седловых задачах оптимизации, а именно на задаче обучения генеративно-состязательных сетей. Также планируется добавить в метод рестарты.

6 Список литературы

[1] Facchinei F., Pang J.S. Finite-Dimensional Variational Inequality and Complementarity Problems. New York: Springer, 2003. V. 1, 2. 693 p

[2] Y. Nesterov, Smooth minimization of non-smooth functions. Math. Program. 103, 127–152 (2005)

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial networks. Commun. ACM 63, 139–144 (2020)

[4] Y. Jin and A. Sidford, Efficiently solving MDPs with stochastic mirror descent. In Proceedings of the 37th International Conference on Machine Learning (ICML), Proc. Mach. Learn. Res. 119, 4890–4900 (2020)

[5] S. Omidshafiei, J. Pazis, C. Amato, J. P. How and J. Vian, Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In Proceedings of the 34th International Conference on Machine Learning (ICML), Proc. Mach. Learn. Res. 70, 2681–2690 (2017)

[6] Корпелевич Г.М. Экстраградиентный метод для отыскания седловых точек и других задач Экономика и матем. методы. Т. 12. № 4. С. 747–756.

[7] Nemirovski A. Prox-method with rate of convergence O(1/T) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems SIAM Journal on Optimization. 2004. V. 15. P. 229–251.

[8] Anton Rodomanov Ali Kavis Yongtao Wu Kimon Antonakopoulos Volkan Cevher Universal Gradient Methods for Stochastic Convex Optimization. 2024.

[9] Bach F., Levy K. Y. A universal algorithm for variational inequalities adaptive to smoothness and noise // arXiv:1902.01637.

[10] Iusem A. N. et al. Variance-based extragradient methods with line search for stochastic variational inequalities // SIAM Journal on Optimization. 2019. V. 29, № 1. С. 175–206.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition // arXiv:1512.03385, 2015

[12] Fedor Stonyakin, Alexander Gasnikov, Pavel Dvurechensky, Alexander Titov, Mohammad Alkousa, Generalized Mirror Prox Algorithm for Monotone Variational Inequalities: Universality and Inexact Oracle // Journal of Optimization Theory and Applications (2022) 194:988–1013

[13] Nemirovski, A.: Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM J. Optim. 15(1), 229–251 (2004)