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 методами на задаче оптимизации для генеративно-состязательных сетей.

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].

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 ограничения константами $D \geq \max\limits_{x,y\in Q} \|x - y\|$, $R^2 \geq \max\limits_{x,y\in Q} V[y](x)$
В стохастическом случаи предполагаем:

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

$$E_\xi \|g(x, \xi) - g(x)\|^2 \leq \sigma^2$$

Ещё будем использовать дивергенцию Брегмана:

$$V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle , \ x, z \in Q$$

Где $d(x)$ - непрерывно дифференцируемая строго выпуклая функция $d : Q \to \mathbb{R}$ и удовлетворяет условию Гёльдера с константой D_μ:

$$\exists \mu \in [0, 1], D_\nu \geq 0 : \ \|\nabla d(x) - \nabla d(y)\|_* \leq D_\mu \|x - y\|^\mu \quad \forall x, y \in Q$$

Тогда $V[z](x) \geq \frac{1}{2}\|x - z\|^2$ и $V[z](x) \leq \frac{D_\mu}{1+\mu}\|x - z\|^{1+\mu}$

3 Универсальный проксимальный зеркальный метод

В проксимальном зеркальном методе мы рассматриваем шаг:

$$w_k = arg\min_{x\in Q}\left(\langle g(z_k), x - z_k\rangle + L_k V[z_k](x)\right)$$

$$z_{k+1} = arg\min_{x\in Q}\left(\langle g(w_k), x - w_k\rangle + L_k V[z_k](x)\right)$$

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

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

$$\langle g(z), z^* - z\rangle + L_k\frac{1}{2}\|z - z^*\|^2 \geq \langle g(w), z^* - w\rangle + L_k\frac{1}{2}\|w - z^*\|^2 + L_k\frac{1}{2}\|z - w\|^2$$

Тогда получим неравенства:

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

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

Преобразуем:

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

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

Введём обозначение:

$$\xi_k[w](z) = \langle g(w), z - w\rangle + L_k\frac{1}{2}\|z - w\|^2$$

Тогда:

$$-\langle g(z_k), z^* - z_k\rangle + L_k\frac{1}{2}\|w_k - z^*\|^2 \leq -\xi_k[z_k](w_k) + L_k\frac{1}{2}\|z_k - z^*\|^2$$

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

Сложим оба неравенства:

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

Из проксимального зеркального метода следует:

$$-\langle g(z_k), w_k - z^*\rangle - \langle g(w_k), z^* - w_k\rangle = \langle g(w_k) - g(z_k), w_k - z^*\rangle \leq L_k\left(\frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}\|w_k - z^*\|^2\right)$$

Снова сложим оба неравенства:

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

$$\leq -\xi_k[z_k](w_k) - \xi_k[w_k](z_{k+1}) + L_k\frac{1}{2}\|z_k - z^*\|^2 + L_k\frac{1}{2}\|w_k - z^*\|^2 + L_k\frac{1}{2}\|z_k - w_k\|^2$$

Получаем:

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

К обоим частям прибавим $L_{k+1}\|z_{k+1} - z^*\|^2$:

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

$$\leq -\xi_k[w_k](z_{k+1}) + (L_{k+1} - \frac{1}{2}L_k)\|z_{k+1} - z^*\|^2 + L_k\frac{1}{2}\|z_k - z^*\|^2 + L_k\frac{1}{2}\|w_k - z^*\|^2$$

Тогда при $L_{k+1} \geq \frac{1}{2}L_k$:

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

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

Будем искать L_{k+1} такое, что:

$$(L_{k+1} - \frac{1}{2}L_k)D^2 = \left|L_k\frac{D^2}{2} - \xi_k[w_k](z_{k+1})\right|_+$$

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

Тогда:

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

Через телескопическую сумму с коэффициентами 2^{i-k} получаем:

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

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

Для доказательства сходимости сделаем оценку на L_{k+1}. Для этого возьмём немного другое L_{k+1} по формуле:

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

Если правая часть отрицательная, то получаем $L_{k+1} = \frac{1}{2}L_k$, иначе:

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

Тогда:

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

где $r = \|w_k - z_{k+1}\|$. Тогда найдём максимум функции $f(r) = L_\nu r^{1+\nu} - L_{k+1}\frac{1}{2}r^2$:

$$f'(r) = (1+\nu)L_\nu r^\nu - L_{k+1}r = 0$$

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

Тогда:

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

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

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

$$(L_{k+1} - L_k)D^2 \leq \frac{L_\nu^{p+1}}{(p+1)((1+\nu)L_{k+1})^p}$$

$$(p+1)L_{k+1}^p(L_{k+1} - L_k) \leq \frac{L_\nu^{p+1}}{D^2(1+\nu)^p} = \alpha$$

Заметим, что:

$$(p+1)L_{k+1}^p(L_{k+1} - L_k) \geq (p+1)\int\limits_{\frac{1}{2}L_k}^{L_{k+1}} t^p dt = L_{k+1}^{p+1} - L_k^{p+1}$$

Тогда:

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

по телескопической сумме:

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

Для достаточно большого k:

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

Тогда:

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

Algorithm 1 Универсальный проксимальный зеркальный метод

1: $Set\ z_0 = arg\min_{u \in Q} d(u),\ L_0 = \|g(z_0)\|, best = z_0$
2: **for** $k = 0, 1, \ldots$ **do**
3: $\quad w_k = arg\min_{x \in Q} \left(\langle g(z_k), x\rangle + L_k V[z_k](x)\right)$
4: $\quad z_{k+1} = arg\min_{x \in Q} \left(\langle g(w_k), x\rangle + L_k V[z_k](x)\right)$
5: $\quad L_{k+1} = \frac{1}{2}L_k + \max\left(0, \frac{1}{2}L_k - \frac{1}{D^2}\langle g(w_k), z_{k+1} - w_k\rangle - L_k \frac{1}{2D^2}\|w_k - z_{k+1}\|^2\right)$
6: **end for**

4 Вычислительный эксперимент

Для сравнения работы универсального градиентного спуска с другими известными методами оптимизации обучим генеративно-состязательную сеть с разными оптимизаторами и построим графики метрик качества предсказаний.

Возьмём датасет образцов рукописного написания цифр MNIST. Он содержит 60000 трейновых и 10000 тестовых картинок размера 28x28, каждая подписана соответствующей ей цифрой.

В качестве модели возьмём простую модель, предложенную в данной статье. В качестве оптимизаторов рассмотрим SGD (Стохастический градиентный спуск), Adam, AdamW (Adam с сокращением веса) и USGM (Универсальный стохастический градиентный спуск). После обучения модели построим графики метрик BCELoss (Бинарный кросс-энтропийный лосс) и WGAN (Wasserstein Loss) [11] зависящих от номера эпохи обучения.

5 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. C. 175–206.

[11] Martin Arjovsky, Soumith Chintala, Léon Bottou, Wasserstein GAN // arXiv:1701.07875, 2017