Стохастическая онлайн оптимизация в случае тяжелых хвостов стохастических градиентов

Гасников А.В.

Кафедра МОУ ФУПМ, ПреМоЛаб МФТИ

gasnikov@yandex.ru

ИОИ-2014, о. Крит (Греция), октябрь, 2014

Рассмотрим задачу выпуклой стохастической оптимизации:

$$f(x) = E_\xi\left[f(x,\xi)\right] \to \min_{x\in Q},$$

где $f(x,\xi)$ – выпуклая по $x \in \mathbb{R}^n$ $(n \gg 1)$ функция. Будем считать, что п.н. $\left\|\nabla f(x,\xi)\right\|_2 \le M$ $(\nabla = \nabla_x$ и E_ξ – перестановочны), а размер выпуклого замкнутого множества Q равен R (в действительности, достаточно считать, что R – расстояние от точки старта до решения, при этом множество Q может быть не ограничено). Можно предложить такой (SA) метод, на каждом шаге (итерации) которого считается проекция стохастического (суб-)градиента функции $f(x,\xi)$ (с независимой разыгранной с.в. ξ) по x на множество Q, что ($\sigma > 0$ – малый доверительный уровень)

$$P_{x_N}\left\{ E_\xi\left[f\left(x_N,\xi\right)\right]-\min_{x\in Q} E_\xi\left[f\left(x,\xi\right)\right]\geq CMR\sqrt{\frac{1+\ln\left(\sigma^{-1}\right)}{N}}\right\}\leq\sigma,$$

где C – константа (~ 10), а с.в. x_N – то, что выдает алгоритм после итераций. Таким образом, для достижения точности по функции ε и доверительного уровня σ методу потребуется $\mathrm{O}\left(M^2R^2\ln\left(\sigma^{-1}\right)/\varepsilon^2\right)$ итераций (вычислений стохастического градиента и его проектирований). Можно показать, что если использовать метод Монте-Карло (SAA), заключающийся в замене исходной задачи следующей задачей

$$\frac{1}{N}\sum_{k=1}^{N} f\left(x,\xi_k\right)\to\min_{x\in Q},$$

где с.в. ξ_k – i.i.d., и распределены также как и ξ, то для того, чтобы гарантировать, что абсолютно точное решение этой новой задачи является (ε,σ)-решением исходной задачи потребуется порядка $\mathrm{O}\left(M^2R^2\left(n\ln\left(MR/\varepsilon\right)+\ln\left(\sigma^{-1}\right)\right)/\varepsilon^2\right)$ итераций.

Приведенный факт хорошо поясняет, что подход, связанный с усреднением случайности за счет самого метода более предпочтителен, чем замена задачи ее стохастической аппроксимацией (эта идея довольно стара; по-видимому, одним из первых кто количественно смог это прочувствовать был Б.Т. Поляк). Более предпочтителен не только тем, что допускает адаптивность постановки и легко переносится на онлайн модификации исходной задачи, но, прежде всего, лучшей приспособленностью к большим размерностям. Подробнее о методах SA написано в статье *Juditsky A.*,

Lan G., Nemirovski A., Shapiro A. Stochastic approximation approach to stochastic programming // SIAM Journal on Optimization. 2009. V. 19. № 4. P. 1574–1609. О методах SAA написано в монографии *Shapiro A., Dentcheva D., Ruszczynski A.* Lecture on stochastic programming. Modeling and theory. MPS-SIAM series on Optimization, 2009. Здесь важно отметить, что в этой задаче “сидит” фундаментальная идея о том, что для получения (агрегирования) хороших оценок неизвестных параметров (особенно когда размерность пространства параметров велика) имеет смысл рассматривать задачу поиска оптимальных значений параметра, как задачу стохастической оптимизации и рассматривать выборку, как источник стохастических градиентов. Например, истинное значение неизвестного вектора параметров в предположении верности исходной параметрической гипотезы может быть записано как решение задачи стохастической оптимизации

$$\theta^* = \arg\max_{\theta \in Q} E_\xi L(\theta, \xi)$$

(метод наибольшего правдоподобия Фишера),

где $L(\theta, \xi)$ – логарифм функции правдоподобия. Однако решать эту задачу обычными методами мы не можем, потому что математическое ожидание берется по с.в. , распределение которой задается $L(\theta^*, \xi)$, а – не известно. Этот порочный круг распутывается, если мы будем решать задачу

$$E_\xi \left[-L(\theta, \xi) \right] \to \min_{\theta \in Q}$$

методами стохастической оптимизации, получая на каждом шаге новую реализацию (элемент выборки) ξ_k и рассчитывая значения градиента $\partial L(\theta, \xi_k)/\partial$. То, что выдает алгоритм и будет (ε, σ)-

оценкой вектора неизвестных параметров. Отметим, что, как правило, дополнительно известно, что $L(\theta,\xi)$ – гладкая и μ-сильно вогнутая (равномерно по ξ) функция по θ. Последнее обстоятельство позволяет получить лучшую скорость сходимости по функции $\mathrm{O}\left(M^2 \ln\left(\ln(N)/\sigma\right)/(\mu N)\right)$, т.е. ($x=\theta$, $f=-L$)

$$P_{x_N}\left\{E_\xi\left[f(x_N,\xi)\right]-\min_{x\in Q}E_\xi\left[f(x,\xi)\right]\geq \bar{C}M^2\frac{\ln\left(\ln(N)/\sigma\right)}{\mu N}\right\}\leq\sigma.$$

Из неравенства Рао–Крамера будет следовать, что такая оценка не улучшаема (с точностью до фактора $\ln\left(\ln(N)\right)$). Правда, тут возникают некоторые тонкости, когда мы говорим о неулучшаемости оценок с учетом вероятностей больших отклонений. Строго говоря, результаты типа Рао–Крамера, Ван-Трисса и т.п. (см., например,

классическую монографию Ибрагимов–Хасьминский) позволяют лишь говорить о неулучшаемости в смысле сходимости полных математических ожиданий (без вероятностей больших отклонений), и именно в таком смысле можно получить неулучшаемую (с точностью до мультипликативной константы) оценку:

$$E\left[f\left(x_N, \xi\right)\right] - \min_{x \in Q} E_{\xi}\left[f\left(x, \xi\right)\right] \leq \frac{\breve{C} M^2}{\mu N}.$$

Можно обобщить рассмотренную задачу на случай, $\left\|\nabla f\left(x, \xi\right)\right\|_2$ когда имеет субгауссовский хвост. Тогда (в том числе в сильно выпуклом случае) вместо $\ln\left(\sigma^{-1}\right)$ стоит писать $\ln^2\left(\sigma^{-1}\right)$. Если же $\left\|\nabla f\left(x, \xi\right)\right\|_2^2$ имеет степенной хвост

$$P\left(\frac{\left\|\nabla f\left(x,\xi\right)\right\|_2^2}{M^2}\geq t\right)=\mathrm{O}\left(\frac{1}{t^{\alpha}}\right),$$

то ($\alpha>2$)

$$P_{x_N}\left(E_{\xi}\left[f\left(x_N,\xi\right)\right]-\min_{x\in Q}E_{\xi}\left[f\left(x,\xi\right)\right]\geq C_{\alpha}MR\frac{\sqrt{N}+\left(N/\sigma\right)^{1/\alpha}}{N}\right)\leq\sigma.$$

Если дополнительно $f\left(x\right)=E_{\xi}\left[f\left(x,\xi\right)\right]$ – μ-сильно выпуклая функция, то ($\alpha>1$)

$$P_{x_N}\left(E_{\xi}\left[f\left(x_N,\xi\right)\right]-\min_{x\in Q}E_{\xi}\left[f\left(x,\xi\right)\right]\geq \bar{C}_{\alpha}M^2\frac{\ln\left(\ln\left(N\right)\right)+\sigma^{-1/\alpha}}{\mu N}\right)\leq\sigma.$$

Если ничего не известно о $\left\|\nabla f\left(x,\xi\right)\right\|_2^2$, кроме $E\left\|\nabla f\left(x,\xi\right)\right\|_2^2 \le M^2$, то по неравенству Маркова (второе неравенство подразумевает μ-сильную выпуклость $f\left(x\right)$)

$$P_{x_N}\left(E_\xi\left[f\left(x_N,\xi\right)\right]-\min_{x\in Q}E_\xi\left[f\left(x,\xi\right)\right]\ge\frac{\widehat{C}MR}{\sigma\sqrt{N}}\right)\le\sigma,$$

$$P_{x_N}\left(E_\xi\left[f\left(x_N,\xi\right)\right]-\min_{x\in Q}E_\xi\left[f\left(x,\xi\right)\right]\ge\frac{\breve{C}M^2}{\sigma\mu N}\right)\le\sigma.$$

Все сказанное выше обобщается и на другие прокс-структуры (не обязательно евклидовы). Так в задачах 8, 9 рассматривается прокс-структура, порожденная “расстоянием” *KL* Кульбака–Лейблера (сильно выпуклом в 1-норме с константой 1 на единичном симплексе – неравенство Пинскера), по-видимому, “наилучшим” образом

подходящая для симплекса (с некоторыми оговорками). Выгода от ее использования в том, что теперь $\left\|\nabla f\left(x,\xi\right)\right\|_{\infty}\le M$, что в типичных ситуациях (см. задачу на теорему Б.С. Кашина о поперечниках) дает оценку константы M в $\sim\sqrt{n}$ раз лучше, а плата за это – увеличение оценки размера области в $\sim\ln n$. Детали имеются в упомянутой в начале замечания статье.

Подчеркнем, что все приведенные здесь оценки, вообще говоря (без дополнительных предположений), неулучшаемы (с точностью до мультипликативных констант). Один пример того, как это можно показать был рассмотрен выше (в общем случае следует смотреть монографию Немировского–Юдина). Сейчас же отметим, что если дополнительно известно, что $f\left(x,\xi\right)$ – гладкая по x функция, с константой Липшица градиента L и(или) сильно выпуклая с

константой μ, а вычисление и проектирование стохастического градиента на каждом шаге находится с неконтролируемой точностью δ, вообще говоря, не случайной природы (в этом месте есть неаккуратность, в действительности, определение δ-оракула, выдающего стохастический градиент, более тонкое[1]), то из недавних результатов Nesterov–Devolder–Glineur и Ghadimi–Lan можно получить такие оценки скорости сходимости (здесь стоит отметить большую работу, проделанную П.Е. Двуреченским, по получению нужного обобщения упомянутых выше результатов)

---

[1] (δ, L)-оракул выдает такие $\left(F(x,\xi), G(x,\xi)\right)$, что $D_\xi\left[G(x,\xi)\right] \le D$, и для любых $x, y \in Q$

$$0 \le E_\xi\left[f(y,\xi)\right] - E_\xi\left[F(x,\xi)\right] - \left\langle E_\xi\left[G(x,\xi)\right], y - x\right\rangle \le \frac{L}{2}\|y-x\|^2 + \delta.$$

$$\min\left\{\mathrm{O}\left(\frac{LR^2}{N^p}+\sqrt{\frac{DR^2}{N}}+N^{p-1}\delta\right),\mathrm{O}\left(LR^2\exp\left(-NC\left(\frac{\mu}{L}\right)^{\frac{1}{p}}\right)+\frac{D}{\mu N}+\left(\frac{L}{\mu}\right)^{\frac{p-1}{p}}\delta\right)\right\},$$

где $C\geq 1$ – некоторая константа, D – дисперсия $\nabla f\left(x,\xi\right)$, а параметр $p\in\left[1,2\right]$ подбирается оптимально исходя из масштаба шума δ. Дисперсию можно уменьшать, запрашивая на одном шаге реализацию стохастического градиента не один раз, а m раз, и заменяя стохастический градиент средним арифметическим, мы уменьшаем дисперсию в m раз. Это имеет смысл делать, если слагаемое, отвечающее стохастичности, доминирует. Важно, что мы при этом не увеличиваем число итераций, и слагаемое $N^{p-1}\delta$ остается прежним. Выписанные оценки характеризуют скорость сходимости в среднем. Они с одной стороны не улучшаемы (причем это остается верно при $\delta=0$ и(или) $D=0$; при $D=0$ мы считаем $\varepsilon\gg n^{-2}$, в противном случае

оценки улучшаемы – методы эллипсоидов и внутренней точки, с оценками числа итераций типа $\sim n^{\alpha}\ln\left(R/\varepsilon\right),\ \alpha\geq 1)$ с точностью до мультипликативной константы, с другой стороны достигаются. В терминах больших отклонений возникает аналогичные оговорки тем, которые мы выше делали с одним исключением – в сильно выпуклом случае появляется дополнительная зависимость от N в множители, содержащем σ (см. выше). Эти результаты переносятся и на прокс-структуры отличные от евклидовой. При этом рассмотрение какой-либо отличной от евклидовой структуры в сильно выпуклом случае (когда минимум достигается на втором выражении), как правило, не имеет смысла, поскольку квадрат евклидовой асферичности p-нормы, возникающий в оценках числа обусловленности прокс-функции в p-норме (это число, в свою очередь, оценивает увеличение числа итераций метода при переходе от евклидовой норме к p-норме),

больше либо равен 1. Равенство достигается на евклидовой норме. Скажем, для 1-нормы эта асферичность оценивается снизу размерностью пространства.

Множество Q должно быть достаточно простой структуры, чтобы на него можно было эффективно проектироваться. Однако в приложениях часто возникают задачи условной минимизации, в которых есть ограничения вида $g(x) \le 0$, где $g(x)$ – выпуклые функции. “Зашивать” эти ограничения в Q, как правило, не представляется возможным в виду вышесказанного требования. Тем не менее, на основе описанного выше можно строить (за дополнительную логарифмическую плату) двухуровневые методы условной оптимизации подобно тем методам, которые описаны в конце главы 2 и 3 монографии *Нестеров Ю.Е.* Методы выпуклой оптимизации. М.: МЦНМО, 2010. При этом на каждом шаге такого

метода потребуется проектироваться на пересечение множества Q с некоторым полиэдром, вообще говоря, зависящим от номера шага.

Все сказанное выше переносится в полной мере на задачи композитной оптимизации (Ю.Е. Нестеров) и частично на монотонные вариационные неравенства (Ю.Е. Нестеров).

Отметим также, что параметры R и μ могут быть не известны априорно или процедуры их оценивания приводят к слишком соответственно завышенным и заниженным результатам. Это может быть проблемой, поскольку знание этих и других параметров требуется методу для расчета величин шагов. Из этой ситуации можно выйти за логарифмическое (по этим параметрам) число рестартов метода. Стартуя, скажем, с и делая, предписанное этим R число шагов, мы проверяем выполняется ли условие ε-близости. Если нет, то полагаем $R := 2R$ и т.д. Это константное время и не отразится на

общих трудозатратах. Аналогичное можно сказать про L и D. Однако если убрать стохастичность, тогда L можно не только эффективнее адаптивно подбирать по ходу самих итераций (увеличив в среднем число итераций не более чем в 4 раза), но и в некотором смысле оптимально самонастраиваться на гладкость функционала на текущем участке пребывания метода (речь идет о недавно предложенном универсальном методе Ю.Е. Нестерова). Для оценки D в ряде случаев бывает эффективнее воспользоваться какой-нибудь статистической процедурой.

В число итераций описанных методов не входит размерность пространства n. Это наталкивает на мысль о возможности использовать эти методы, например, в гильбертовых пространствах. Оказывается, это, действительно, можно делать. В частности, концепция неточного оракула позволяет привнести сюда элемент

новизны, существенно мотивированной практическими нуждами (много материала о градиентных методах решения задач оптимизации в гильбертовых пространствах собрано во втором томе учебника Ф.П. Васильева по методам оптимизации).

Наконец, полезно иметь в виду, что за счет допускаемой неточности оракула, выдающего (стохастический) (суб-)градиент, можно “погрузить” задачу с гельдеровым градиентом

$$\left\|\nabla f\left(x,\xi\right)-\nabla f\left(y,\xi\right)\right\|_{*}\leq L_{\nu}\left\|x-y\right\|^{\nu}$$

(в том числе и негладкую задачу с ограниченной нормой разности субградиентов $\nu=0$) в класс гладких задач с оракулом, характеризующимся точностью δ и

$$L = L_\nu \left[\frac{L_\nu (1-\nu)}{2\delta(1+\nu)} \right]^{\frac{1-\nu}{1+\nu}}.$$

В стохастической онлайн ситуации оценки будут такими:

$$\min\left\{ \mathrm{O}\left(\sqrt{\frac{M^2 R^2}{N}} + \delta \right), \mathrm{O}\left(\frac{M^2 \ln N}{\mu N} + \delta \right) \right\}.$$

Эти оценки достигаются и неулучшаемы. Как видно из этих оценок наличие гладкости ничего не дает. Все что ранее говорилось про прокс-структуру и большие отклонения полностью и практически без изменений (в μ-сильно выпуклом случае в оценках вероятностей больших уклонений $\ln\left(\ln\left(N\right)\right) \Rightarrow \ln N$) переносится и на онлайн случай. Отметим также, что онлайн методы, как правило, допускают

прямо-двойственную модификацию (с теми же оценками скорости сходимости) при применении к задачам условной оптимизации.

В заключение отметим, что следует различать задачи стохастической оптимизации и задачи, в которые мы сами искусственно привносим случайность (рандомизацию) с целью более эффективного решения задачи. К этой ситуации, скажем, можно отнести случай, когда (негладкий) выпуклый функционал в задаче детерминированный, но представляет собой трудно вычислимый интеграл от параметров, который компактно представим в виде математического ожидания по некоторой не сложной вероятностной мере. Тогда выгоднее считать стохастический градиент. Существенно экономя на вычислениях на каждом шаге и лишь не много теряя на логарифмическом увеличении числа шагов. Такой пример будет рассмотрен в следующей задаче. Однако если мы можем вычислять

значение функции в детерминированной задаче, в которую мы решали рандомизированным методом, то ни о каких тяжелых хвостах можно не заботиться. Поскольку, выбрав число шагов так, чтобы метод находил ε-решение с вероятностью $\geq 1/2$ и запустив реализаций такого метода мы за дополнительную $\log_2\left(\sigma^{-1}\right)$ плату (мультипликативную) получим с вероятностью $1-\sigma$ среди выданных ответов хотя бы одно ε-решение. Однако предположение о возможности вычислять значения функции в ряде задач “натыкается” на существенные вычислительные сложности (значительно большие, чем при расчете стохастического градиента). Таким примером является задача поиска вектора PageRank $P^T p = p$ (P– стохастическая матрица), сводящаяся к негладкой задаче выпуклой оптимизации (матричной игре) $\max\limits_{u \in S_n(1)} \left\langle u, P^T p - p \right\rangle \to \min\limits_{p \in S_n(1)}$. Кроме того, если

рандомизация осуществляется каким-то специальным образом, например, таким, что

$$E\left[\left\|\nabla f\left(x,\xi\right)\right\|_*^2\right] \le C_n \left\|\nabla f\left(x\right)\right\|_*^2 + \textit{малый добавок}\left(\varepsilon\right)$$

и в точке минимума $\nabla f\left(x\right)$, то приведенные выше оценки можно существенно улучшить (Б.Т. Поляк). Такая рандомизация возникает, например, в связи с изучением безградиентных методов.