

© 2015 г. А.В. ГАСНИКОВ, к.ф.-м.н.

(ПреМоЛаб МФТИ, ИППИ РАН, ВШЭ)

П.Е. ДВУРЕЧЕНСКИЙ, к.ф.-м.н.

(ПреМоЛаб МФТИ, ИППИ РАН, WIAS Berlin)

Ю.Е. НЕСТЕРОВ, д.ф.-м.н.

(CORE UCL Louvain la Neuve, Belgium)

Стохастические градиентные методы с неточным оракулом

В работе предпринята попытка описать современное состояние проксимальных градиентных методов (в том числе прямых методов и методов покомпонентного спуска) решения задач выпуклой стохастической оптимизации с неточным оракулом (неточность неслучайной природы), выдающим стохастический субградиент. Заметная часть приведенных в статье результатов была получена (авторами статьи и их коллегами) относительно недавно. Цель данной работы – собрать все воедино и посмотреть на разнообразные факты из этой области с единой позиции.

1. Введение

В 1960-е годы численные методы выпуклой оптимизации переживали свою первую большую революцию. В работах того времени четко и последовательно развивалась линия градиентных методов. Основополагающим здесь можно признать вклад Бориса Теодоровича Поляка [1, 2], с работ которого во многом и началось активное и повсеместное использование градиентных методов в Советском Союзе. Следующая революция началась в конце 1970-х годов после фундаментальных работ А.С. Немировского, Д.Б. Юдина, Л.Г. Хачияна, N. Karmarkar'a и др. [3, 4]. В монографии [4] была предложена классификация задач выпуклой (и не только) оптимизации по степени гладкости и выпуклости. Были получены нижние оценки для соответствующих классов задач оптимизации с оракулом, выдающим по запросу градиент или стохастический градиент, его компоненту или просто значение функции в точке. Стало понятно, чего в

принципе можно достичь. Стали строиться оптимальные методы, см., например, [5–7]. При этом на задачи стали смотреть более пристально с точки зрения теории сложности. Появилась битовая сложность. Была показана полиномиальная разрешимость задач линейного программирования в битовой сложности [3]. Началась разработка полиномиальных методов внутренней точки для задач выпуклой оптимизации на базе метода Ньютона, которая впоследствии привела к созданию общей теории [8–11] и соответствующего пакета CVX [12], способного решать широкий спектр задач выпуклой оптимизации в пространствах размерности до $n \sim 10^4 - 10^5$. Однако, вызовы нового тысячелетия заставляют снова вернуться к градиентным методам. Задачи, которые стали возникать в последние десять лет, отличаются огромными размерностями $n \sim 10^6 - 10^9$. Такие задачи (классифицируемые как задачи large-scale и huge-scale оптимизации) приходят из анализа данных, поиска равновесий в различных сетевых моделях (связанных с компьютерными и транспортными сетями), биоинформатики и многих других областей. Для таких размерностей шаг (итерация) метода Ньютона, становится слишком дорогим, поэтому приходится снова возвращаться к более медленным (в смысле скорости сходимости), но более дешевым (в смысле стоимости одной итерации) градиентным методам. Но для указанных размерностей даже градиентные методы могут испытывать проблемы. В этой связи оказалась очень полезной концепция “заглядывания в черный ящик”, т.е. в структуру задачи с целью ускорения вычислений [13], и использование вместо градиента его легко вычислимой (стохастической) аппроксимации [11]. Как следствие, стало принято считать, что правильный способ эффективно решать ту или иную задачу – это отказаться от общих методов, оптимальных на больших классах, и погружаться в специфику конкретной задачи в надежде ускориться и получить оценки лучше, чем нижние границы [4]. Можно сказать, что началась новая революция. Поток работ на эту тему в основных профилирующих журналах (например, Math. Program.) резко возрос (см., например, обзор [11]). Тем не менее, параллельно стали появляться работы (в том числе работы авторов статьи), показывающие, что многие эффективные методы решения современных задач выпуклой оптимизации в пространствах огромных размеров получаются сочетанием небольшого количества приемов и идей. Цель настоящей работы состоит в том, чтобы собрать воедино набор основных таких идей и показать их связь с некоторыми концепциями 1960-х годов, многие из которых восходят к Б.Т. Поляку. Мы сосредоточимся на оценках числа итераций, требующихся различным методам для решения задачи выпуклой оптимизации с заданной точностью по функции. Эта информация не в полной мере характеризует эффективность метода, но она необходима для последующего его полного исследования. Мы также ограничимся рассмотрением, так называемых, проксимальных градиентных методов (методов проекции градиента) [14, 15], в которые, например, не входят очень популярные в последнее время методы условного градиента [11, 16, 17]. В качестве основного инструментария для получения эффективных методов используется метод оценивающих последовательностей, восходящий к работам одного из авторов статьи [7, 10, 13]. Здесь имеются и альтернативные подходы, например, [18–20]. Из-за ограничений на объем статьи и большого количества технических деталей

мы ограничимся здесь лишь изложением общей картины. В частности, в статье не приводится псевдокод соответствующих алгоритмов, но, как правило, указываются источники, в которых его можно найти. Мы также не претендуем здесь на полный обзор современного состояния исследований, посвященных проксимальным градиентным методам. Более того, при ссылках на литературу мы далеко не всегда ссылались на первоисточники, иногда предпочитая ссыльаться на удачно написанный более доступный и более современный обзор или монографию.

2. Стохастическая оптимизация

Рассматривается задача выпуклой стохастической оптимизации [2, 21]:

$$(1) \quad f(x) = E_{\xi} [f(x, \xi)] \rightarrow \min_{x \in Q},$$

где $f(x)$ – выпуклая по $x \in \mathbb{R}^n$ ($n \gg 1$) функция. Будем называть $\nabla f(x, \xi)$ стохастическим субградиентом функции $f(x, \xi)$ в точке x по первой переменной. Будем считать, что¹ п.н. $\|\nabla f(x, \xi)\|_2 \leq M$, $\nabla = \nabla_x$ и E_{ξ} – перестановочны.² Обозначим через R – диаметр выпуклого замкнутого множества Q : $R = \max_{x, y \in Q} \|x - y\|_2$. В действительности, достаточно считать, что R – расстояние от точки старта до решения. При этом множество Q может быть (с некоторыми оговорками, см. п. 3) не ограничено [24]. Мы будем считать, что множество Q простой структуры, т.е. на него можно эффективно проектироваться. В работах [24–28] рассматривались различные варианты методов проекции градиента с усреднением и длинными шагами³ применительно к решению задачи (1). Общая оценка скорости сходимости этих методов есть ($\sigma > 0$ – малый доверительный уровень)⁴

¹ В действительности [22], здесь и практически в любом другом контексте, где возникает такого типа условия, достаточно требовать, что п.н. выполнено неравенство $\|\nabla f(y, \xi) - \nabla f(x, \xi)\|_2 \leq M$. Это позволяет в ряде случаев понизить оценку константы $\nabla f(x, \xi)$ и как следствие (см. (2)), ускорить метод. Отметим, что под $\nabla f(x, \xi)$ (аналогично под $\nabla f(y, \xi)$) понимается любой элемент соответствующего субградиента.

² Для задач онлайн оптимизации условие перестановочности необходимо записывать в более общем (маргинальном) виде [23].

³ Б.Т. Поляком было показано [25], что такое “робастное” сочетание позволяет получать эффективные методы для данного класса задач.

⁴ Эта оценка неулучшаема с точностью до мультипликативной константы C (даже в детерминированном случае $f(x, \xi) \equiv f(x)$), см. [4].

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

где C – константа (здесь и далее константы в основном будут в диапазоне $\sim 10^0 - 10^2$), а случайный вектор x_N – то, что выдает алгоритм (например, метод зеркального спуска [27] или метод двойственных усреднений [24] – сравнительный анализ и описание “физики” этих методов в детерминированном случае проводится в работе [20]) после N итераций. Мы будем называть x_N – (ε, σ) -решением задачи (1). Таким образом, для достижения точности по функции ε и доверительного уровня σ методу потребуется (здесь и далее мы будем использовать $O(\cdot)$, однако все эти формулы могут быть переписаны с точными константами, что важно, поскольку во многих ситуациях такие оценки используются в для формирования критерия останова метода)

$$(2) \quad O(M^2 R^2 \ln(\sigma^{-1}) / \varepsilon^2)$$

итераций (вычислений стохастического субградиента и проектирований).

Отметим, что если использовать метод Монте-Карло, заключающийся в замене исходной задачи следующей задачей

$$(3) \quad \frac{1}{N} \sum_{k=1}^N f(x, \xi_k) \rightarrow \min_{x \in Q},$$

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 взять N порядка [29]

$$O(M^2 R^2 (n \ln(MR/\varepsilon) + \ln(\sigma^{-1})) / \varepsilon^2).$$

Эта наблюдение хорошо поясняет, что подход, связанный с усреднением случайности за счет самого метода более предпочтителен, чем замена задачи (1) ее стохастической аппроксимацией (3).⁵ Более предпочтителен не только тем, что допускает адаптивность

⁵ Особенno ярко это проявляется в случае, бесконечномерных пространств, возникающих в статистической теории обучения (СТО = SLT, Statistical Learning Theory) [30]. Попытка обучиться за счет минимизации эмпирического риска (а именно так можно расшифровать формулу (3) в СТО) может не дать состоятельной

постановки и легко переносится на онлайн модификации исходной задачи, но, прежде всего, лучшей приспособленностью к большим размерностям.

Здесь важно подчеркнуть фундаментальную идею⁶, которую можно усмотреть, например, в цикле работ Б.Т. Поляка с Я.З. Цыпкиным [31], 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 [32, 33] (метод наибольшего правдоподобия Фишера)

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

где $L(\theta, \xi)$ – логарифм функции правдоподобия. Однако решать эту задачу обычными методами мы не можем, потому что математическое ожидание берется по с.в. ξ , распределение которой задается неизвестным параметром θ^* . Обойти эту сложность можно, если решать ту же самую задачу

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

методами стохастической оптимизации, получая на каждом шаге новую реализацию (элемент выборки) ξ_k и рассчитывая значения стохастического градиента $\partial L(\theta, \xi_k)/\partial \theta$.

То, что выдает алгоритм и будет оценкой вектора неизвестных параметров θ^* . Как правило, дополнительно известно, что $L(\theta, \xi)$ – гладкая и μ -сильно вогнутая (равномерно по ξ) функция от θ . Последнее обстоятельство позволяет получить лучшую оценку скорости сходимости по функции [11, 34–36] (в [11, 34] используются специальная

оценки/решающего правила, в то время как соответствующий стохастический зеркальный спуск дает состоятельную оценку. Отметим, что в работе [30] приводится достаточно интересный общий результат: в задачах обучения (в частности, в задачах СТО, математической статистики и онлайн обучения) способ получения оптимальных (с точностью до логарифмических факторов) оценок/решающих правил (или, другими словами, способ наискорейшего обучения) базируется на применении соответствующего метода зеркального спуска. Правда, найти “соответствующий метод”, в свою очередь, представляет собой не простую задачу.

⁶ Распространяемую и на непараметрическую статистику. Отметим, что начиная с 1980-х годов XX века в этом направлении был цикл работ А.С. Немировского, Б.Т. Поляка и А.Б. Цыбакова, оказавших заметное влияние и на текущие исследования в этой области.

модификация метода проекции градиента с усреднением и выбором шагов $h_k = 2(\mu k)^{-1}$ и $h_k = (\mu k)^{-1}$, о подходе [35] и близком ему подходе [36] будет не много написано в п. 3)

$$(4) \quad O(M^2 \ln(\ln(N)/\sigma)/(\mu N)),$$

т.е. ($x = \theta$, $f = -L$),

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

Из неравенства Рао–Крамера [33] будет следовать, что оценка (4) – не улучшаемая (с точностью до фактора $\ln(\ln(N))$). Правда, тут возникают некоторые тонкости, когда мы говорим о неулучшаемости оценок с учетом вероятностей больших отклонений. Строго говоря, классические результаты типа Рао–Крамера, Ван–Трисса и т.п. (см., например, [33]) позволяют лишь говорить о неулучшаемости в смысле сходимости полных математических ожиданий (без вероятностей больших отклонений), и именно в таком смысле можно получить (с помощью методов [11, 34–36]) неулучшаемую (с точностью до мультипликативной константы) оценку:

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

Можно обобщить рассмотренную постановку задачи (1) на случай, когда $\|\nabla f(x, \xi)\|_2$ имеет субгауссовский хвост (определение см., например, в [27]). Тогда (в том числе в сильно выпуклом случае) вместо $\ln(\sigma^{-1})$ стоит писать $\ln^2(\sigma^{-1})$. Если же $\|\nabla f(x, \xi)\|_2^2$ имеет степенной хвост [37]

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

где $\alpha > 2$, то⁷

⁷ Приводимые ниже неравенства стоит понимать так, что x_N выдается методом [24, 27], а в сильно выпуклом случае, методом [34–36]. При этом для оценок вероятностей больших отклонений в случае тяжелых хвостов требуется некоторые оговорки и уточнения. К сожалению, мы не смогли найти соответствующий выписанным оценкам (в случае тяжелых хвостов) источник литературы. Приведенные здесь нами формулы нуждаются в дополнительной апробации.

$$P_{x_N} \left(f(x_N) - \min_{x \in Q} f(x) \geq C_\alpha M R \frac{\sqrt{N} + (N/\sigma)^{1/\alpha}}{N} \right) \leq \sigma.$$

Если дополнительно $f(x) = E_\xi [f(x, \xi)]$ — μ -сильно выпуклая функция, то (при $\alpha > 1$)

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

Если ничего не известно о $\|\nabla f(x, \xi)\|_2^2$, кроме неравенства $E_\xi [\|\nabla f(x, \xi)\|_2^2] \leq M^2$, то по неравенству Маркова

$$P_{x_N} \left(f(x_N) - \min_{x \in Q} f(x) \geq \frac{\hat{C} M R}{\sigma \sqrt{N}} \right) \leq \sigma,$$

$$P_{x_N} \left(f(x_N) - \min_{x \in Q} f(x) \geq \frac{\check{C} M^2}{\sigma \mu N} \right) \leq \sigma,$$

второе неравенство подразумевает μ -сильную выпуклость $f(x)$.

Можно задать вопрос: насколько вообще уместно рассматривать постановки, в которых возникают тяжелые хвосты. Ведь, если мы можем эффективно вычислять значения функции $f(x) = E_\xi [f(x, \xi)]$ в задаче (1), 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 $\log_2(\sigma^{-1})$ реализаций такого метода мы за дополнительную $\log_2(\sigma^{-1})$ плату (мультиплекативную) получим с вероятностью $1 - \sigma$ среди выданных ответов хотя бы одно ε -решение [28, 36]. Однако предположение о возможности эффективно вычислять значения функции (при условии трудной вычислимости ее градиента), как правило, не встречается на практике. В некотором смысле типичным тут является пример 1 (см. ниже) вычисления вектора PageRank (при $n \sim 10^9$). Собственно, все это (сложность в расчете значения функции, если градиент тяжело считается) неплохо соответствует философии быстрого автоматического дифференцирования (БАД) [38, 39]. Согласно теории БАД, если мы можем посчитать значение функции, то мы можем не более чем в 4 раза дороже посчитать и ее градиент. Как следствие, если мы можем эффективно посчитать значение $f(x)$, то, как правило, мы и $\nabla f(x)$ можем эффективно посчитать. Тогда и на исходную задачу (1) можно смотреть уже не как на задачу стохастической оптимизации, а как на обычную задачу выпуклой оптимизации, что может существенно ускорить ее решение (см. п. 3 ниже).

Отметим также (следуя А.С. Немировскому), что с помощью концепции неточного оракула (см. п. 3 ниже) мы можем редуцировать задачу с тяжелыми хвостами $\|\nabla f(x, \xi)\|_2^2$ к ситуации, когда п.н. $\|\nabla f(x, \xi)\|_2 \leq M(\varepsilon)$. Для этого нужно “обрезать” стохастический градиент

$$\nabla f(x, \xi) := \begin{cases} \nabla f(x, \xi), & \|\nabla f(x, \xi)\|_2 \leq M(\varepsilon) \\ M(\varepsilon) \frac{\nabla f(x, \xi)}{\|\nabla f(x, \xi)\|_2}, & \|\nabla f(x, \xi)\|_2 > M(\varepsilon) \end{cases}.$$

Константа $M(\varepsilon)$ подбирается оптимальным образом, исходя из желаемой точности ε . Чем больше $M(\varepsilon)$, тем меньше смещение (bias) обрезанного стохастического градиента, как следствие, тем точнее можно восстановить решение исходной задачи, но при этом возрастает необходимое число итераций (см. (2), (4)). Оптимальный выбор этой константы (с точностью до логарифмического фактора) дает приведенные выше оценки.

Все сказанное выше⁸ обобщается и на другие прокс-структуры [4] (не обязательно евклидовы, когда выбирается прокс-функция $d(x) = \|x\|_2^2$), согласно которым осуществляется (как правило, по явным формулам) “проектирование” на Q . Так для множества $Q = S_n(1)$ ($S_n(R) = \left\{x \in \mathbb{R}^n : x_i \geq 0, i = 1, \dots, n, \sum_{i=1}^n x_i = R\right\}$ – единичный симплекс в n -мерном пространстве) часто рассматривается (см. пример 1 вычисления вектора PageRank ниже) KL-прокс-структура: $d(x) = \ln n + \sum_{i=1}^n x_i \ln x_i$. Эта прокс-функция $d(x)$ сильно выпукла в 1-норме с константой сильной выпуклости $\alpha = 1$ на $S_n(1)$ – неравенство Пинскера [24, 27]. Она “наилучшим” образом подходит для симплекса (с некоторыми оговорками [23, 40]). Выгода от ее использования в том, что норма стохастического субградиента всегда оценивается в сопряженном пространстве к пространству, в котором прокс-функция 1-сильно выпукла. В рассматриваемом случае получается $\|\nabla f(x, \xi)\|_\infty \leq M$, что в типичных ситуациях дает оценку константы M в $\sim \sqrt{n}$ раз лучше, чем 2-норме, а плата за это – увеличение оценки размера области ($R^2 = \max_{x \in Q} d(x)/\alpha$) в $\sim \ln n$ раз. Детали имеются, например, в статье [27]. Здесь мы отметим (следуя А.С. Немировскому), что в общем случае оптимальный (с некоторыми

⁸ В сильно выпуклом случае можно ограничиться 2-нормой, порождающей евклидову прокс-структуру. Попытка выбрать какую-либо отличную прокс-структуру в сильно выпуклом случае, как правило, не только не улучшает оценку, но и портит её (см. п. 3 и [35]).

оговорками – см. ниже) выбор прокс-структурой связан с симметризацией множества Q : $B = (Q - Q)/2$. Выпуклое центрально симметричное множество B порождает по теореме Колмогорова норму, в которой B является единичным шаром. Далее ищется оптимальная прокс-функция, согласованная с этой нормой. Говоря более формально, ищется такая сильно выпуклая в этой норме функция $d(x) \geq 0$ с константой сильной выпуклости $\alpha \geq 1$, чтобы $R^2 = \max_{x \in Q} d(x)/\alpha$ было бы минимально возможным. Если $Q = B_2^n(1)$ – единичный евклидов шар, то $R^2 = 1$, т.е. не зависит от размерности пространства n , но если $Q = B_\infty^n(1)$ – единичный шар в l_∞^n норме, то $R^2 = \Omega(n)$ (т.е. существует такое χ , что при достаточно больших значениях n имеет место неравенство $R^2 \geq \chi n$). Как будет видно из замечания 2 (на примере когда $Q = B_\infty^n(1)$), такой выбор нормы не всегда приводит к оптимальным во всех смыслах оценкам (аналогичные примеры нам встретятся и в следующих двух пунктах).

Замечание 1. Стоит обратить внимание на то, что если выбрана евклидова прокс-структура, то R^2 – квадрат евклидова диаметра Q . При переходе к другой прокс-структуре в качестве R^2 фигурирует прокс-диаметр Q ($\text{diam}(Q) = \max_{x \in Q} d(x)$), поделенный на константу сильной выпуклости $\alpha = \alpha(Q)$ прокс-функции, заданной на Q , относительно выбранной нормы в прямом пространстве. Скажем, в случае выбора KL-прокс-структурь, 1-нормы в прямом пространстве и $Q = S_n(r)$, имеем

$$R^2 = \text{diam}(S_n(r))/\alpha(S_n(r)) = r \cdot \text{diam}(S_n(1))/(\alpha(S_n(1))/r) = r^2 \cdot (\ln n)/1 = r^2 \ln n.$$

Для евклидовой прокс-структурь размер $Q = S_n(r)$ равнялся бы $2r^2$. Отсюда можно сделать вывод (верный и в общем случае), что выбор прокс-структурь имеет целью оптимально учесть структуру множества с точки зрения того как в итоговую оценку числа итераций будет входить размерность пространства, в котором происходит оптимизация. При гомотетичном увеличении/уменьшении множества оценки числа итераций будут меняться одинаково, независимо от выбранной прокс-структурь. Отметим также, что в формуле (2) для прокс-структурь, отличной от евклидовой точнее писать не $R^2 \ln(\sigma^{-1})$, где $R^2 = r^2 \ln n$ (приводим для KL-прокс-структурь), а $r^2(\ln n + \ln(\sigma^{-1})) = r^2 \ln(n/\sigma)$.

Замечание 2. Пусть $Q = B_q^n(1)$ – единичный шар в q -норме или, в более общем случае, Q содержится в $B_q^n(1)$. Относительно оптимального выбора нормы и прокс-структурь можно заметить следующее (см., например, [4, 41]): если $q \geq 2$, то в качестве нормы оптимально выбирать $\|\cdot\|_2$ (2-норму) и евклидову прокс-структурь. Определим q' из $1/q + 1/q' = 1$. Пусть $1 \leq q \leq 2$, тогда $q' \geq 2$. Если при этом $q' = o(\log n)$, то оптимально выбирать $\|\cdot\|_q$, а прокс-структурь задавать прокс-функцией $d(x) = \frac{1}{q-1} \|x\|_q^2$. Во всех этих случаях $R^2 = O(1)$. Для $q' \geq \Omega(\log n)$, выберем $a = 2\log n/(2\log n - 1)$,

$\| \cdot \| = \| \cdot \|_a$, а прокс-структуру будем задавать прокс-функцией $d(x) = \frac{1}{a-1} \|x\|_a^2$. В этом случае $R^2 = O(\log n)$.

Не сложно проверить, что для единичного симплекса, вложимого в единичный шар в 1-норме, выбор соответствующих прокс-структур из замечаний 1, 2 приводит к одинаковому результату в категориях $O(\cdot)$. В частности, для случая когда $Q = B_\infty^n(1)$, выбор 2-нормы и евклидовой прокс-структуры приводит к оценке (напомним, что речь идет только об оценках типа (2), для оценок типа (4) прокс-структура всегда евклидова) $O(M_2^2 n \ln(\sigma^{-1})/\varepsilon^2)$ вместо $O(M_\infty^2 n \ln(\sigma^{-1})/\varepsilon^2)$, получаемой при выборе l_∞^n нормы в прямом пространстве. Аналогично вышеннаписанному можно отметить, что в типичных ситуациях последняя оценка может быть в $\sim n$ раз хуже. Тем не менее, оценка $O(M_\infty^2 n \ln(\sigma^{-1})/\varepsilon^2)$ не улучшаема в общем случае. Потому что в общем случае нет гарантий, что $M_2^2 \ll M_\infty^2$, а если такие гарантии есть, то это уже сужает класс функций, для которого получена нижняя оценка с константой M_∞^2 .

Подчеркнем, что приведенные здесь оценки (2), (4) (в детерминированном случае при дополнительном условии, что требуемое число итераций для достижения точности ε удовлетворяет неравенству $N(\varepsilon) \leq n$ [4]) без дополнительных предположений являются неулучшаемыми (с точностью до мультипликативных констант) для класса задач стохастической оптимизации (1) и негладких детерминированных задач. Причем дополнительная гладкость функционала задачи (1) в стохастической постановке в общем случае не приводит к улучшению приведенных оценок (2), (4). Если делать дополнительные предположения о малости случайного шума (low noise conditions), то приведенные оценки можно улучшать (см. п. 3). Один пример того, как можно устанавливать неулучшаемость оценок был рассмотрен выше (на основе неравенства Рао-Крамера), в общем случае следует смотреть монографию [4] и [41]. Отметим, что в работе [41] показывается, что для задач стохастической оптимизации (1) при оптимизации на шарах в q -норме оценки типа (2), даваемые методами зеркального спуска с прокс-структурой из замечания 2, соответствуют с точностью до логарифмического фактора нижним оценкам.

Следует, однако, различать задачи стохастической оптимизации и задачи, в которые мы сами искусственно привносим случайность (используя рандомизацию) с целью уменьшения числа арифметических операций на одну итерацию метода [28]. К последнему можно отнести случай, когда (негладкий) выпуклый функционал в задаче является детерминированным, но представляет собой трудно вычислимый интеграл (сумму), зависящую от (оптимизируемых) параметров, который может быть компактно представлен в виде математического ожидания по некоторой простой вероятностной мере. Тогда выгоднее вычислять на каждой итерации метода стохастический градиент, существенно экономя на вычислениях на каждом шаге и лишь немногого теряя на логарифмическом увеличении числа шагов ($\sim \ln(\sigma^{-1})$). Ярким примером на эту тему является Google problem (PageRank). По-видимому, одними из первых на эту задачу

посмотрели в указанном выше контексте А.В. Назин и Б.Т. Поляк в работе [42], см. также [43–45].

Пример 1 (PageRank). Задача поиска вектора PageRank p из уравнения $P^T p = p$ (P – стохастическая матрица), сводится [44, 45] к негладкой задаче выпуклой оптимизации (седловой задаче)

$$\max_{u \in S_n(1)} \langle u, P^T p - p \rangle \rightarrow \min_{p \in S_n(1)}.$$

Перепишем эту задачу в общем виде

$$\min_{x \in S_n(1)} \max_{y \in S_n(1)} \langle y, Ax \rangle,$$

где матрица A большого размера $n \times n$ (вообще говоря, не разреженная) с элементами, ограниченными по модулю числом $M = 1$. Ключевое наблюдение для решения этой задачи состоит в том [27, 42], что:

$$Ax = E_{i[x]} \left[A^{\langle i[x] \rangle} \right],$$

где $A^{\langle i \rangle}$ – i -й столбец матрицы A , вектор $x \in S_n(1)$, а с.в. $i[x]$ имеет мультиномиальное распределение с вектором параметров x . Важным следствием является тот факт, что левая часть равенства, Ax , вычисляется за $O(n^2)$ арифметических операций, а выражение, стоящее в правой части под математическим ожиданием, $A^{\langle i[x] \rangle}$ – всего лишь за $O(n)$ арифметических операций. Используя это наблюдение (и аналогичное для умножения матрицы A на вектор-строку слева), можно показать, что (рандомизированный) метод зеркального спуска [27] (с KL-прокс-структурой) и стохастическим градиентом по x равным $A^{\langle i[x] \rangle}$ (аналогично по y) после

$$O\left(\frac{nM^2 \ln(n/\sigma)}{\varepsilon^2}\right) = O\left(\frac{n \ln(n/\sigma)}{\varepsilon^2}\right)$$

элементарных арифметических операций выдает такие $x \in S_n(1)$ и $y \in S_m(1)$, что

$$\max_{\tilde{y} \in S_m(1)} \tilde{y}^T Ax - \min_{\tilde{x} \in S_n(1)} y^T A \tilde{x} \leq \varepsilon$$

с вероятностью $\geq 1 - \sigma$.

Если дополнительно известно, что матрица P – разрежена, то можно организовать поиск (ε, σ) -решения еще эффективнее – рандомизировать при проектировании на

симплекс [23, 45]. Тогда вместо фактора n в оценках общего числа операций будет фигурировать “среднее” число элементов матрицы P (по строкам и столбцам) отличных от нуля.

Отметим, что в определенных ситуациях (например, при условии $n \gg \varepsilon^{-2}$ – типичном для задач huge-scale оптимизации) такому рандомизированному методу потребуется считать относительно небольшое количество элементов матрицы A за все время работы, в то время как для класса детерминированных алгоритмов потребуется считать как минимум половину элементов матрицы A [3] для $\varepsilon = 0.1$.

Хочется также отметить, что на задаче из примера 1 можно продемонстрировать большую часть современного инструментария, необходимого для решения задач huge-scale оптимизации. Так, в случае разреженной матрицы A для решения поставленной негладкой задачи выпуклой оптимизации (и многих других) хорошо подходит метод Б.Т. Поляка [2, 44], работающий по нижним оценкам (2) (функционал негладкий) и при этом учитывающий разреженность A при пересчете градиента [44]. Другой подход [43] (задача поиска вектора PageRank сводится к минимизации другого функционала), также нашедший широкое применение [46], связан с заменой градиентного спуска на покомпонентный спуск. Такая замена увеличивает в среднем число итераций в n раз (см. п. 4), но зато (благодаря разреженности) происходит экономия при пересчете одной компоненты градиента, большая, чем в n раз по сравнению с расчетом полного градиента. Поясним это следующим примером.

Пример 2 (разреженный PageRank). Задача поиска вектора PageRank также может быть сведена к следующей задачи выпуклой оптимизации

$$f(x) = \frac{1}{2} \|Ax\|_2^2 + \frac{1}{2} \sum_{k=1}^n (-x_k)_+^2 \rightarrow \min_{\langle x, e \rangle = 1}$$

где как и в примере 1 $A = P^T - I$, I – единичная матрица, $e = (1, \dots, 1)^T$,

$$(-y)_+ = \begin{cases} y, & y \geq 0 \\ 0, & y < 0 \end{cases}.$$

При этом мы считаем, что в каждом столбце и каждой строке матрицы P не более $s \ll \sqrt{n}$ элементов отлично от нуля (P – разрежена). Этую задачу предлагается решать обычным градиентным методом, но не в евклидовой норме, а в 1-норме (см., например, [20]):

$$x_{k+1} = x_k + \arg \min_{h: \langle h, e \rangle = 0} \left\{ f(x_k) + \langle \nabla f(x_k), h \rangle + \frac{L}{2} \|h\|_1^2 \right\},$$

где $L = \max_{i=1,\dots,n} \|A^{(i)}\|_2^2 + 1 = O(1)$. Для достижения точности ε^2 по функции потребуется сделать $O(LR^2/\varepsilon^2) = O(1/\varepsilon^2)$ итераций [13]. Не сложно проверить, что пересчет градиента на каждой итерации заключается в умножении $A^T Ah$, что может быть сделано за $O(s^2)$. Связано это с тем, что вектор h всегда имеет только две компоненты отличные от нуля (такая разреженность получилась благодаря выбору 1-нормы), причем эти компоненты соответствуют $\arg \min_{i=1,\dots,n} \partial f(x_k)/\partial x^i$ и $\arg \max_{i=1,\dots,n} \partial f(x_k)/\partial x^i$, что пересчитывается (при использовании специального двоичного дерева для поддержания максимальной и минимальной компоненты градиента [44]) за $O(s^2 \ln n)$. Таким образом, общая трудоемкость предложенного метода будет $O(s^2 \ln n/\varepsilon^2)$, что заметно лучше многих известных методов [45]. Стоит также отметить, что функционал, выбранный в этом примере, обеспечивает намного лучшую оценку $\|Ax\|_2 \leq \varepsilon$ по сравнению с функционалом из примера 1, который, с некоторой натяжкой, обеспечивает $\|Ax\|_\infty \leq \varepsilon$. Наилучшая (в разреженном случае без условий на спектральную щель матрицы P [45]) из известных нам на данный момент оценок $O(s \ln(n/\sigma)/\varepsilon^2)$ [23, 45] для $\|Ax\|_\infty$ может быть улучшена приведенной в этом примере оценкой, поскольку, как уже отмечалось ранее, $\|Ax\|_2$ может быть (и так часто бывает) в $\sim \sqrt{n}$ раз больше $\|Ax\|_\infty$, а $s \ll \sqrt{n}$. Не большая “ложка дегтя” в том, что в решении могут быть маленькие отрицательные компоненты.

Мы привели здесь этот пример, с одной стороны, потому что он демонстрирует определенный антагонизм (вместе с примером 1) рандомизации и разреженности: как правило, удается сполна воспользоваться чем-то одним (упомянутый пример из [23, 45] пока единственное известное нам исключение). С другой стороны, пример 2 также характерным образом демонстрирует каким образом используется разреженность (см. также [44]). Обратим внимание на то, что число элементов в матрице P отличных от нуля, даже при наложенном условии разреженности (по строкам и столбцам), все равно может быть достаточно большим sn . Удивляет то, что в оценке общей трудоемкости этот размер фактически никак не присутствует. Это в перспективе (при правильной организации работы с памятью) позволяет решать задачи колоссальных размеров. Более того, даже в случае не большого числа не разреженных ограничений вида $\langle a_i, x \rangle = b_i$, $i = 1, \dots, m = O(1)$, можно раздуть пространство (не более чем в два раза), в котором происходит оптимизация (во многих методах, которые учитывают разреженность, это число входит под логарифмом подобно рассмотренному примеру, так что такое раздутье не приведет к серьезным затратам), и переписать эту систему в виде $Ax = b$, где матрица будет иметь размеры $O(n) \times O(n)$, но число отличных от нуля элементов в каждой строке

и столбце будет $O(1)$. Таким образом, допускается небольшое число “плотных” ограничений.

Приведем еще один пример, подсказывающий, как следует решать задачу (3), полученную из (1) с применением идеи метода Монте-Карло.

Пример 3 (рандомизация суммы). Пусть необходимо решить задачу выпуклой оптимизации

$$(5) \quad f(x) = \frac{1}{N} \sum_{k=1}^N f_k(x) \rightarrow \min_{x \in Q},$$

где $f_k(x)$ – негладкие выпуклые функции с ограниченной числом M нормой субградиента, Q – выпуклое замкнутое множество простой структуры (можем эффективно на него проектироваться, согласно заданной прокс-функции) проск-диаметра R . Введем новую функцию

$$f(x, \xi) = \begin{cases} f_1(x), & \text{с вероятностью } 1/N \\ \dots \\ f_N(x), & \text{с вероятностью } 1/N \end{cases}.$$

Ее стохастический субградиент легко вычислить. Для этого разыгрывается за $O(\ln N)$ с.в. ξ , принимающая значения $1, \dots, N$ с равными вероятностями (см., например, [45]). Затем считается субградиент $f_\xi(x)$ (и выполняется прокс-проектирование на Q). Как уже отмечалось ранее, можно найти (ε, σ) -решение так понимаемой задачи (5) за $O(M^2 R^2 \ln(\sigma^{-1})/\varepsilon^2)$ итераций, со стоимостью одной итерации равной $O(\ln N)$ + затраты на вычисления субградиента $f_\xi(x)$ + затраты на вычисление проекции. Если решать задачу без рандомизации, то число итераций будет $O(M^2 R^2/\varepsilon^2)$, строго говоря, здесь M должно быть немного меньше за счет того, что

$$\max_{x \in Q} \left\| \frac{1}{N} \sum_{k=1}^N \nabla f_k(x) \right\|_* \leq \max_{\substack{k=1, \dots, N \\ x \in Q}} \|\nabla f_k(x)\|_*,$$

но мы считаем, что обе части неравенства одного порядка. Зато шаг итерации будет теперь почти в N раз дороже. И если $N \gg 1$ это может оказаться существенным.

Приведенную постановку можно распространить на случай, когда взвешивание функций не равномерное (тогда первое разыгрывание/приготовление займет $O(N)$, а все

последующие $O(\ln N)$) и $f_k(x) := E_{\xi_k} [f_k(x, \xi_k)]$ с равномерно ограниченными (по k , x и ξ) нормами субградиентов. При этом все приведенные оценки числа итераций сохраняются. Причем требование равномерной ограниченности норм субградиентов можно существенно ослабить за небольшую плату (см. выше).

Если на решение задачи (3) теперь посмотреть в контексте описанной рандомизации с $f_k(x) = f(x, \xi_k)$ (здесь ξ_k – не случайная величина, а полученная в методе Монте Карло k -я по порядку реализация с.в. ξ), то “все встанет на свои места” в смысле одинаковости (с точностью до логарифмического фактора) двух подходов к решению задачи (1), описанных в начале пункта.

Описанная рандомизация при вычислении субградиента суммы функций, по-видимому, была одной из первых, которые предлагались в стохастической оптимизации [21]. Однако она популярна и по сей день, например, в связи с приложениями к поиску равновесий в транспортных сетях [47] и анализу данных [48]. В частности, в [11, 49] в предположении, что все функции в (5) гладкие с константой Липшица градиента L , предложен специальный рандомизированный метод, в котором число вычислений градиентов

$$O\left((N + LR^2/\varepsilon)(\ln(LR^2/\varepsilon) + \ln(\sigma^{-1}))\right),$$

а если дополнительно имеется еще и μ -сильная выпукłość, то (считаем $N \gg \sqrt{L/\mu}$)

$$O\left((N + L/\mu)(\ln(LR^2/\varepsilon) + \ln(\sigma^{-1}))\right).$$

Отметим, что вторая оценка переходит в первую при следующей квадратичной регуляризации. К выпуклому функционалу прибавляется регуляризирующее слагаемое $\mu \|x\|_2^2$. В результате функционал становится сильно выпуклым и справедлива вторая оценка на число вычислений градиента. Такая регуляризация изменяет исходную целевую функцию на число не больше μR^2 и чтобы итоговая погрешность по исходной функции была порядка ε нужно выбирать $\mu \sim \varepsilon/R^2$. При подстановке этого значения во вторую оценку числа вычислений градиента последняя переходит в первую оценку.

Отметим также, что сначала в [49] получается результат о сходимости средних

$$E\left(f(x_N) - \min_{x \in Q} f(x)\right) \leq \varepsilon,$$

где

$$N = N(\varepsilon) = O\left((N + L/\mu)\ln(LR^2/\varepsilon)\right),$$

Потом из неравенства Маркова получают оценку больших уклонений

$$P\left(f(x_{N(\varepsilon)}) - \min_{x \in Q} f(x) \geq \sigma\right) \leq \varepsilon/\sigma,$$

которую переписывают в виде

$$P\left(f(x_{N(\varepsilon\sigma)}) - \min_{x \in Q} f(x) \geq \sigma\right) \leq \varepsilon,$$

где

$$N(\varepsilon\sigma) = O\left((N + L/\mu)(\ln(LR^2/\varepsilon) + \ln(\sigma^{-1}))\right).$$

Мы привели здесь это наблюдение, потому что оно оказывается полезным и во многих других контекстах, в которых рандомизированный метод сходится со скоростью геометрической прогрессии.

Так, в книге [2] Б.Т. Поляк отмечает, что если рандомизация осуществляется каким-то специальным образом, например, таким, что

$$(6) \quad E\left[\|\nabla f(x, \xi)\|_*^2\right] \leq C_n \|\nabla f(x)\|_*^2 + \Delta,$$

где $\Delta \geq 0$ некоторая малая погрешность, и в точке минимума $\nabla f(x) = 0$, то приведенные выше оценки (2), (4) можно существенно улучшить. Примеры будут приведены ниже в п.4 (см. (22)). В частности, в сильно выпуклом случае можно получить геометрическую скорость сходимости. Важно отметить, что при рандомизации, возникающей в покомпонентных спусках, спусках по направлению и безградиентных методах в гладком случае условие (6) выполняется [2, 43, 50]. Мы вернемся к этому кругу вопросов в п. 4. Описанная же выше конструкция (с довольно грубым неравенством Маркова) используется в данном контексте [43, 50] для (точной!) оценки больших уклонений. Причем за счет регуляризации функционала, о которой было сказано выше, все это переносится и просто на гладкий случай без предположения сильной выпуклости.

3. Стохастические градиентные методы с неточным оракулом

В данном пункте мы опишем что можно получить, если дополнительно известно, что $f(x)$ – гладкая по x функция, с константой Липшица градиента L и(или) сильно

выпуклая с константой $\mu \geq 0$, но вычисление стохастического градиента на каждом шаге происходит с неконтролируемой неточностью δ , вообще говоря, не случайной природы.⁹

Замечание 3. И гладкости и сильной выпуклости можно добиться искусственно. Как уже отмечалось в п. 2, сильная выпуклость всегда легко получается регуляризацией функционала в исходной задаче. Как правило, это не дает ничего нового с точки зрения выписанных оценок (и даже ухудшает эти оценки на логарифмический фактор), но в ряде специальных случаев, когда у задачи, есть дополнительная (например, седловая) структура, это может давать определенные преимущества. Кроме того, такая регуляризация иногда просто необходима для корректности постановки. Это связано с тем, что в общем случае даже для гладких детерминированных выпуклых задач мы можем гарантировать сходимость итерационного метода лишь по функции, но не по аргументу. Для сходимости по аргументу нужна сильная выпуклость функционала, которую и обеспечивают должной регуляризацией (см., например, конец п. 2). Идея регуляризации используется в популярном методе двойственного сглаживания [51] (регуляризация двойственной задачи с целью улучшения гладких свойств прямой). В отличие от прямой регуляризации, эта техника хорошо работает только для вполне конкретных задач, имеющих определенную структуру, когда исходная задача имеет явное двойственное представление, введя в которое регуляризацию, можно явно пересчитать во что превратится исходная прямая задача. Другой пример сглаживания будет приведен в п. 4.

Сформулируем более точно предположения об оракуле, выдающем стохастический градиент, следуя [52, 53].¹⁰

Предположение 1. (δ, L) -оракул выдает (на запрос, в котором указывается только одна точка x) такие $(F(x, \xi), G(x, \xi))$ (с.в. ξ независимо разыгрывается из одного и того же распределения, фигурирующего в постановке (1)), что для всех $x \in Q$ ограничена дисперсия

$$E_\xi \left[\left\| G(x, \xi) - E_\xi [G(x, \xi)] \right\|_*^2 \right] \leq D,$$

и для любых $x, y \in Q$

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

⁹ Особое внимание таким постановкам стали уделять после выхода книг [2, 31], одним из авторов, которых является Б.Т. Поляк. В них обстоятельно изучается “влияние помех”, в том числе не случайной природы, на методы выпуклой оптимизации.

¹⁰ В работе [52] собрано много различных мотиваций такому предположению (определению). В определенном смысле это предположение 1 наиболее общее и, одновременно, наиболее точно отражающее спектр всевозможных приложений.

Из недавних результатов [52–60] можно получить общий метод (мы приводим огрубленный вариант оценки времени работы этого метода для большей наглядности), с такими оценками скорости сходимости¹¹

$$(7) \quad \min \left\{ O\left(\frac{LR^2}{N^{p+1}} + \sqrt{\frac{DR^2}{N}} + N^p \delta \right), O\left(LR^2 \exp\left(-\Upsilon N \cdot \left(\frac{\mu}{L}\right)^{\frac{1}{p+1}} \right) + \frac{D}{\mu N} + \left(\frac{L}{\mu}\right)^{\frac{p}{p+1}} \delta \right) \right\},$$

где $\Upsilon \geq 1$ – некоторая константа, а параметр $p \in [0,1]$ подбирается оптимально перед запуском метода исходя из масштаба шума δ . Для лучшего понимания оценки (7) бывает полезно ее переписать в еще более огрубленном виде

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

Этот общий метод – есть в некотором смысле “выпуклая комбинация” двойственного градиентного метода (DGM) и быстрого градиентного метода (FGM) [57, 60], оценки скорости сходимости для которых имеют соответственно вид:

$$(DGM) \quad \min \left\{ O\left(\frac{LR^2}{N} + \sqrt{\frac{DR^2}{N}} + \delta \right), O\left(LR^2 \exp\left(-\Upsilon_1 N \frac{\mu}{L} \right) + \frac{D}{\mu N} + \delta \right) \right\},$$

$$(FGM) \quad \min \left\{ O\left(\frac{LR^2}{N^2} + \sqrt{\frac{DR^2}{N}} + N \delta \right), O\left(LR^2 \exp\left(-\Upsilon_2 N \sqrt{\frac{\mu}{L}} \right) + \frac{D}{\mu N} + \sqrt{\frac{L}{\mu}} \delta \right) \right\}.$$

Комбинируя эти два метода можно непрерывно настраиваться (оптимально подбирая метод, регулируя $p \in [0,1]$) на шум (известного масштаба). В этой связи также полезно отметить, что FGM есть специальная выпуклая комбинация прямого градиентного метода (PGM), оценки скорости сходимости которого совпадают с оценками DGM, и метода зеркального спуска / двойственных усреднений [20] (по-видимому, здесь вместо зеркального спуска можно использовать и метод из работы [61]). Нельзя в этой связи не обратить внимание на то, что комбинация двух методов привела к новому методу, работающему лучше чем каждый из методов по отдельности.

¹¹ Оценки характеризуют достигнутую в среднем точность (по оптимизируемому функционалу) после N итераций. При этом в случае когда минимум достигается на втором аргументе (выгодно использовать факт наличия μ -сильной выпуклости) под N правильнее понимать не число итераций, а число обращений к (δ, L) -оракулу [60].

Вся последующая часть п. 3 будет посвящена обсуждению этих результатов и их окрестностей.

Прежде всего заметим, что дисперсию у первого аргумента минимума в (7) можно уменьшать в m раз, запрашивая на одном шаге реализацию стохастического градиента не один раз, а m раз, и заменяя стохастический градиент средним арифметическим [11, 28, 53] (в случае тяжелых хвостов у стохастических градиентов лучше пользоваться более робастными оценками, например, медианного типа [4]). Это имеет смысл делать, если слагаемое, отвечающее стохастичности, доминирует. Важно, что мы при этом не увеличиваем число итераций, и слагаемое $N^p\delta$ остается прежним. Отметим, что число вызовов оракула при этом увеличивается, но тем не менее, в некоторых ситуациях такой подход может оказаться оправданным. Такая игра используется¹² в способе получения второго аргумента оценки (7). В этой связи оценку (7) правильнее переписать следующим образом (здесь $N(\varepsilon)$ – число обращений к (δ, L) -оракулу, необходимых для достижения в среднем по функции точности ε , индекс 1 соответствует просто выпуклому, а индекс 2 сильно выпуклому случаю):

$$(8) \quad N_1(\varepsilon) = \max \left\{ O\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p+1}}, O\left(\frac{DR^2}{\varepsilon^2}\right) \right\}, \quad N_2(\varepsilon) = \max \left\{ O\left(\left(\frac{L}{\mu}\right)^{\frac{1}{p+1}} \ln\left(\frac{LR^2}{\varepsilon}\right)\right), O\left(\frac{D}{\mu\varepsilon}\right) \right\}$$

при (условия на допустимый уровень шума, при котором оценки (8) имеют такой же вид, с точностью до $O(1)$, как если бы шума не было)

$$(9) \quad \delta_1(\varepsilon) \leq O\left(\varepsilon \cdot \left(\frac{\varepsilon}{LR^2}\right)^{\frac{p}{p+1}}\right), \quad \delta_2(\varepsilon) \leq O\left(\varepsilon \cdot \left(\frac{\mu}{L}\right)^{\frac{p}{p+1}}\right).$$

Как уже отмечалось, выписанные оценки (7) ((8), (9)) характеризуют скорость сходимости в среднем. Они с одной стороны не улучшаются¹³ с точностью до мультипликативной константы (см. п. 2 и [4, 41]), а с другой стороны достигаются. Все

¹² Вместе с идеей рестартов [35, 58–60], распространяющей (ускоряющей) практически любой итерационный метод (с явной оценкой необходимого числа итераций $N(\varepsilon)$ для достижения заданной точности ε) на случай сильно выпуклого функционала. Нетривиально здесь то, что при довольно общих условиях при таком распространении сохраняется (и работает уже в условиях сильной выпуклости) свойство оптимальности исходного метода.

¹³ В нижнюю оценку во втором выражении под знаком минимума при экспоненте вместо LR^2 входит μR^2 , а константа $\Upsilon=1$ в (7). Впрочем, получить вместо фактора LR^2 фактор μR^2 можно аккуратно проанализировав оценки, даваемые с помощью техники рестартов (см., например, [20, 60]).

это (неулучшаемость оценок) справедливо и при $\delta = 0$ и(или) $D = 0$. При этом в случае $D = 0$, необходимо считать, что требуемое число итераций для достижения точности ε удовлетворяет неравенству $N(\varepsilon) \leq n$ [4], в противном случае оценки улучшаемы – методы эллипсоидов и внутренней точки [4, 8–11], с оценками числа итераций типа $O(n^\alpha \ln(BR/\varepsilon))$, $\alpha \geq 1$, $|f(x)| \leq B$. В терминах больших отклонений возникают оценки, аналогичные тем, которые были приведены в п. 2, см. [60].

Замечание 4. Отметим, что пока нам не известно строгое обоснование оценок (7) ((8), (9)) с вероятностями больших отклонений для случая не ограниченного множества Q . В известном нам способе получения оценок вероятностей больших уклонений, к сожалению, явно используется предположение об ограниченности множества Q . Это ограничение, скорее всего, может быть опущено в коэрцитивном случае (диаметр, равно как и размер R , считаются в норме, которая выбрана в прямом пространстве, – прокс-диаметр здесь не нужен)

$$\text{diam}\{x \in Q : f(x) \leq f(x_0)\} = O(R),$$

где x_0 – точка старта метода. Связано это с тем, что все рассматриваемые методы обладают тем свойством [20, 24], что стартуя из R окрестности решения, они не уйдут за пределы его $O(R)$ окрестности по ходу всего движения. Это, в свою очередь, следует из того, что мы строим все методы на основе двух базисных PGM и методе зеркального спуска / двойственных усреднений (со схожими свойствами с DGM). Коэрцитивность нужна для PGM [20]. Требование коэрцитивности не является сильно обременительным. Его можно обеспечивать за счет регуляризации задачи [62]. Тонкость тут еще и в том, что требуется оговорка “не уйдут с очень большой вероятностью”.

Отмеченные результаты переносятся и на прокс-структуры отличные от евклидовой [60]. При этом рассмотрение какой-либо другой q -нормы (l_q -нормы) в прямом пространстве ($q \geq 1$), отличной от евклидовой, в сильно выпуклом случае (когда минимум достигается на втором выражении в (7)), как правило, не имеет смысла. Связано это с тем, что квадрат евклидовой асферичности q -нормы, возникающий в оценках числа обусловленности прокс-функции в q -норме (это число, в свою очередь, оценивает увеличение числа итераций метода при переходе от евклидовой норме к q -норме), больше либо равен 1. Равенство достигается на евклидовой норме. Скажем, для 1-нормы эта асферичность оценивается снизу размерностью пространства [13, 35].

Полезно также иметь в виду, что за счет допускаемой неточности оракула, можно погрузить задачу с гельдеровым градиентом $\|\nabla f(x) - \nabla f(y)\|_* \leq L_\nu \|x - y\|^\nu$, $\nu \in [0, 1]$ (в том числе и негладкую задачу с ограниченной нормой разности субградиентов при $\nu = 0$) в класс гладких задач с оракулом, характеризующимся точностью δ и [52]

$$(10) \quad L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)} \right]^{\frac{1-\nu}{1+\nu}}.$$

Заметим, в этой связи, что если в предположении 1 считать

$$E_\xi[f(y, \xi)] - E_\xi[F(x, \xi)] - \langle E_\xi[G(x, \xi)], y - x \rangle \leq \frac{L}{2} \|y - x\|^2 + M \|y - x\| + \delta,$$

то вместо D в (7) ((8), (9)) стоит писать $M^2 + D$.

Таким образом, например, можно получить оценки (2), (4) из оценки (7). Но наряду с введенной нами игрой на точности оракула, исходный оракул может еще иметь свою неточность. Несложно привести оценки (на базе формулы (7) и ((8), (9))) сочетающие все это вместе.

Множество Q должно быть достаточно простой структуры, чтобы на него можно было эффективно проектироваться. Однако в приложениях часто возникают задачи условной минимизации [2], в которых, например, есть ограничения вида $g(x) \leq 0$, где $g(x)$ – выпуклые функции [14]. Зашивать эти ограничения в Q , как правило, не представляется возможным ввиду вышесказанного требования о легкости проектирования. Тем не менее, на основе описанного выше можно строить (за дополнительную логарифмическую плату) двухуровневые методы (наверное, лучше говорить “методы уровней”, чтобы не было путаницы с многоуровневой оптимизацией, см. пример 5) условной оптимизации [9, 10, 63]. При этом на каждом шаге такого метода потребуется проектироваться на пересечение множества Q с некоторым полиэдром, вообще говоря, зависящим от номера шага. Последнее обстоятельство в общем случае сужает класс задач, к которым применимы такие многоуровневые методы до класса задач, к которым применимы методы внутренней точки [10]. В частности, возникает довольно обременительное условие на размер пространства, в котором проходит оптимизация: $n \sim 10^4 - 10^5$. Все это не удивительно, поскольку имеются нижние оценки [4] (рассматриваются аффинные ограничения в виде равенств, аналогично могут быть рассмотрены и неравенства), показывающие, что в общем случае для нахождения такого $x \in \mathbb{R}^n$, что $\|Ax - b\|_2 \leq \varepsilon$ потребуется не меньше, чем $\Theta(\sqrt{L_x R_x} / \varepsilon)$ операций типа умножения Ax ($L_x = \sigma_{\max}(A) = \lambda_{\max}(A^T A)$, $R_x = \|x^*\|_2 = \|A^T (AA^T)^{-1} b\|_2$). Аналогичное можно сказать и про седловые задачи: для отыскания такой пары (x, y) , что (левая часть этого неравенства всегда неотрицательная)

$$\max_{\tilde{y} \in S_n(1)} \tilde{y}^T Ax - \min_{\tilde{x} \in S_n(1)} y^T A \tilde{x} \leq \varepsilon$$

потребуется не меньше, чем $\Theta(\Lambda / \varepsilon)$ (Λ – максимальный по модулю элемент матрицы A) операций типа умножения Ax и $y^T A$. Заметим, что обе выписанные нижние оценки

справедливы при условии, что число итераций (операций типа умножения Ax) $k \leq n$. Как следствие, в общем случае даже для гладкой детерминированной сильно выпуклой постановки при наличии всего лишь аффинных ограничений $Ax = b$ нельзя надеяться на быстрое решение. Тем не менее, некоторые дополнительные предположения в ряде случаев позволяют ускорить решение таких задач (см., например, [62]).

Замечание 5. Задача поиска такого x^* , что $Ax^* = b$ сводится к задаче выпуклой гладкой оптимизации $f(x) = \|Ax - b\|_2^2 \rightarrow \min_x$. Нижние оценки для скорости решения такой задачи [4] (см. также формулу (7) с $\delta = D = 0$, $p = 1$) приводят к оценке: $f(x_k) \geq \Omega(L_x R_x^2 / k^2)$. Откуда следует, что только при $k \geq \Omega(\sqrt{L_x} R_x / \varepsilon)$ можно гарантировать выполнение неравенства $f(x_k) \leq \varepsilon^2$, т.е. $\|Ax_k - b\|_2 \leq \varepsilon$.

Заметим, что эта нижняя оценка для специальных матриц может быть улучшена. Причем речь идет не о недавних результатах D. Spielman'a [64] (премия Неванлины 2010), а о более простой ситуации. Вернемся к задаче поиска вектора PageRank (примеры 1, 2), которую мы перепишем как

$$Ax = \begin{pmatrix} P^T - I \\ 1 \dots 1 \end{pmatrix} x = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = b, \quad I - \text{единичная матрица.}$$

По теореме Фробениуса–Перрона [65] решение такой системы с неразложимой матрицей P единственно и положительно $x > 0$. Сведем решение этой системы уравнений к вырожденной задаче выпуклой оптимизации $\frac{1}{2} \|x\|_2^2 \rightarrow \min_{Ax=b}$. Построим двойственную к ней задачу [9]

$$\min_{Ax=b} \frac{1}{2} \|x\|_2^2 = \min_x \max_{\lambda} \left\{ \frac{1}{2} \|x\|_2^2 + \langle b - Ax, \lambda \rangle \right\} = \max_{\lambda} \min_x \left\{ \frac{1}{2} \|x\|_2^2 + \langle b - Ax, \lambda \rangle \right\} = \max_{\lambda} \left\{ \langle b, \lambda \rangle - \frac{1}{2} \|A^T \lambda\|_2^2 \right\}.$$

Зная решение $\lambda^* = (AA^T)^{-1}b$ двойственной задачи $\langle b, \lambda \rangle - \frac{1}{2} \|A^T \lambda\|_2^2 \rightarrow \max_{\lambda}$ можно восстановить решение прямой задачи (формула Демьянова–Данскина о дифференцировании максимума [9]): $x(\lambda) = A^T \lambda$. Однако важно здесь то, что FGM [10] для этой двойственной задачи, ввиду линейности градиента двойственного функционала по λ , дает возможность попутно получать следующую оценку на норму этого градиента:

$$\|Ax_k - b\|_2 = O\left(\frac{L_y R_y}{k^2}\right),$$

где x_k есть известная выпуклая комбинация $\{x(\lambda_i)\}_{i=1}^k$, $L_y = \sigma_{\max}(A^T)$, $R_y = \|\lambda^*\|_2 = \|(AA^T)^{-1}b\|_2$. Кажется, что это противоречит нижней оценке $\|Ax_k - b\|_2 \geq \Omega(\sqrt{L_x} R_x / k)$. Однако, важно напомнить [4], что эта нижняя оценка установлена для всех $k \leq n$ (n – размерность вектора x), и она будет улучшена, в результате описанной процедуры только если дополнительно предположить, что матрица A удовлетворяет следующему условию $L_y R_y \ll n \sqrt{L_x} R_x$, что сужает класс, на котором была получена нижняя оценка $\Omega(\sqrt{L_x} R_x / k)$. Поскольку

$$R_x^2 = \left\langle (AA^T)^{-1} b, b \right\rangle, R_y^2 = \left\langle (AA^T)^{-1} b, (AA^T)^{-1} b \right\rangle,$$

то в типичных ситуациях можно ожидать $R_y \gg R_x$, что мешает выполнению требуемого условия.

Пример 4. Если имеется дополнительная информация о структуре седловой задачи, то можно её использовать для ускорения [13, 66]. Более того, многие современные постановки задач (негладкой) выпуклой оптимизации (в частности, связанные с compressed sensing и l_1 -оптимизацией) в пространствах огромных размеров специально стараются представить седловым образом с целью получения эффективного решения (см. работы А.С. Немировского, А.Б. Юдицкого [28, 67, 68]). Далее будет разобран один простой пример (немного обобщающий результаты [52, 53]), демонстрирующий возможности градиентных методов с неточным оракулом в седловом контексте. Рассматривается седловая задача ($x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$)

$$f(x) = \max_{\|y\|_1 \leq R_y} \{G(y) + \langle By, x \rangle\} \rightarrow \min_{\|x\|_1 \leq R_x},$$

где функция $G(y)$ – сильно вогнутая с константой κ относительно 1-нормы и константой Липшица градиента L_G (также в 1-норме)¹⁴, а максимальный по модулю элемент матрицы B ограничен числом Λ . Тогда функция $f(x)$ будет гладкой, с константой Липшица градиента $L_f = \Lambda^2/\kappa$. Казалось бы, что мы можем решить задачу минимизации функции $f(x)$ за $O(\sqrt{\Lambda^2 R_x^2 \ln n / (\kappa \varepsilon)})$ итераций. Но это возможно, только если мы можем абсолютно точно находить $\nabla f(x) = By^*(x)$, где $y^*(x)$ – решение вспомогательной задачи максимизации по y . В действительности, мы можем решать эту задачу (при различных x) лишь приближенно. Если мы решаем вспомогательную задачу быстрым градиентным методом [13] с точностью $\delta/2$ (на это потребуется $O(\sqrt{L_G/\kappa} \ln(L_G R_y^2 \ln m / \delta))$ итераций), то выдаваемая по формуле $\nabla f(x) = By_{\delta/2}(x)$ аппроксимация градиента будет $(\delta, 2L_f)$ -оракулом [52, 53]. Выбирая $\delta = O(\varepsilon \sqrt{\varepsilon / (L_f R_x^2 \ln n)})$ (см. формулу (9) при $p=1$), получим после

¹⁴ К сожалению, в 1-норме всегда имеет место неравенство $L_G/\kappa \geq n$ [35]. Если рассматривать в условиях седловой задачи в примере 4 вместо шаров в 1-норме шары в 2-норме (евклидовы), то таких проблем можно избежать (и писать вместо $R_x^2 \ln n$ и $R_y^2 \ln m$ просто R_x^2 и R_y^2), но приобрести вместо Λ^2 потенциально большее число $\sigma_{\max}(B)$ [51].

$$O\left(\sqrt{\frac{L_G \Lambda^2 R_x^2 \ln n}{\kappa^2 \varepsilon}} \ln \left(\frac{L_f L_G R_x^2 R_y^2 \ln n \ln m}{\varepsilon} \right) \right)$$

итераций (на итерациях производится умножение матрицы B на вектор/строчку и вычисление градиента $G(y)$) ε -решение задачи минимизации $f(x)$. Отметим, что если не использовать сильную вогнутость функции $G(y)$, то для получения оценки

$$\max_{\|y\|_1 \leq R_y} \{G(y) + \langle By, x_N \rangle\} - \min_{\|x\|_1 \leq R_x} \{G(y_N) + \langle By_N, x \rangle\} \leq \varepsilon$$

потребуется $\Theta\left(\max\{L_G R_y^2 \ln m, \Lambda R_x R_y \sqrt{\ln n \ln m}\}/\varepsilon\right)$ итераций (см., например, [4, 9, 11]).

Интересно отдельно разобрать ситуацию, когда вместо множества $\|y\|_1 \leq R_y$ фигурирует симплекс $S_m(R_y)$, а $G(y) = -\sum_{k=1}^m y_k \ln(y_k/R_y)$ – сильно вогнутая в 1-норме с константой $\kappa=1$, $R_x = \infty$ (энтропийно-линейное программирование [62]). В этом случае мы не можем обеспечить даже равномерной ограниченности градиента функции $G(y)$. Тем не менее, также можно рассчитывать [62] на зависимость $O(\varepsilon^{-1/2} \ln(\varepsilon^{-1}))$ числа итераций от точности ε для критерия:

$$\max_{y \in S_m(R_y)} \min_x \{G(y) + \langle By, x \rangle\} - \min_x \{G(y_N) + \langle By_N, x \rangle\} \leq \varepsilon.$$

При этом вместо энтропии в качестве функции $G(y)$ можно брать любую сильно вогнутую в 1-норме функцию, для которой решение задачи максимизации (вычисление $f(x)$ с точностью ε) может быть осуществлено за $O(\ln(\varepsilon^{-1}))$. В примере с энтропией, для $f(x)$ есть просто явная формула. Точнее, важно то, что есть явная формула для оптимального решения $y^*(x)$. К сожалению, имеется проблема вхождения в оценку необходимого числа итераций неизвестного размера решения задачи минимизации $f(x)$. Это проблема решаема [62]. В частности, в случае, когда $G(y)$ имеет ограниченную вариацию на множестве $S_m(R_y)$ (для энтропии эта вариация равна $R_y \ln m$), можно предложить метод, с оценкой числа итераций $O(\varepsilon^{-1} \ln(\varepsilon^{-1}))$. В эту оценку уже никак не входит неизвестный размер решения x , который может оказаться большим [62]. Далее мы еще вернемся к вопросу о том, как действовать, в случае, когда тот или иной параметр задачи (в данном случае размер решения) априорно не известен.

Пример 4 был приведен, прежде всего, потому, что он поясняет одно интересное и достаточно современное направление в численных методах выпуклой оптимизации (см., например, [63, 67, 68]). Грубо говоря, это направление можно охарактеризовать, как попытку ввести оптимальную “алгебру” над алгоритмами выпуклой оптимизации. А именно, если требуется оптимизировать функционал (искать седловую точку), который обладает разными свойствами (гладкости, сильной выпуклости, быстроты вычислимости частных производных и т.п.) по разным группам переменным (такие задачи часто в последнее время возникают в разных приложениях, в частности, в транспортных и экономических [47, 69–72]) и(или) сам представляет собой некоторую суперпозицию других функционалов (с разными свойствами; наиболее популярен случай суммы двух функционалов [13, 48, 63, 67, 68]), то хотелось бы получить такую декомпозицию исходной задачи, чтобы правильное сочетание (правильное чередование с правильными частотами) оптимальных методов для получившихся отдельных подзадач позволило бы получить оптимальный метод для исходной задачи. В ряде интересных случаев такое оказывается возможным (с оговоркой, что оптимальность понимается с точностью до логарифмического фактора). По-видимому, новым в этом направлении является наблюдение, отмеченное в примере 4, что при определенных условиях идея оптимального сочетания различных методов для решения одной сложной по структуре задачи оптимизации, может быть реализована на основе концепции неточного оракула.

Другой способ борьбы с дополнительными ограничениями в задачах выпуклой оптимизации, который дает даже больше, базируется на прямо-двойственной структуре [24] всех обсуждаемых методов (поскольку они строят модель функции [13]). Это означает, что ограничения вносятся во вспомогательную задачу оптимизации, возникающую на каждом шаге метода и отвечающую за проектирование. В результате на каждом шаге решается более сложная задача. Тем не менее, если такие вспомогательные задачи можно эффективно решать (что в общем случае также наталкивается на сложности, описанные ранее) с помощью метода множителей Лагранжа (найдя и сами множители), то описанные методы позволяют не только эффективно решать исходную задачу оптимизации с ограничениями, но и находить попутно (по явно выписываемым формулам) решение двойственной задачи.

Основная идея работы [24] состоит в том (здесь мы ограничимся рассмотрением детерминированного случая), что метод генерирует в прямом пространстве на итерациях такую последовательность $\{x_k\}$,¹⁵ что зазор двойственности (duality gap) $\Delta(\lambda; N)$ удовлетворяет условию

¹⁵ В двойственном пространстве при этом генерируется последовательность соответствующих множителей Лагранжа $\{y_k\}$ [24] или, в случае наличия модели у исходной прямой задачи, последовательность $\{y_k\}$ генерируется по явным формулам $\{y_k = y(x_k)\}$ согласно этой модели [17].

$$\Delta(\lambda; N) = \max_{u \in Q} \left\{ \frac{1}{S_k} \sum_{k=1}^N \lambda_k \langle \nabla f(x_k), x_k - u \rangle \right\} \leq \varepsilon,$$

где $S_k = \sum_{k=1}^N \lambda_k$, $\lambda_k \geq 0$, поэтому

$$f\left(\frac{1}{S_k} \sum_{k=1}^N \lambda_k x_k\right) - \min_{x \in Q} f(x) \leq \varepsilon.$$

Это следует из выкладки

$$f\left(\frac{1}{S_k} \sum_{k=1}^N \lambda_k x_k\right) - f(u) \leq \frac{1}{S_k} \sum_{k=1}^N \lambda_k \cdot (f(x_k) - f(u)) \leq \frac{1}{S_k} \sum_{k=1}^N \lambda_k \langle \nabla f(x_k), x_k - u \rangle.$$

Аналогичную точность (для двойственной задачи) дает следующая аппроксимация решения двойственной задачи

$$y = \frac{1}{S_k} \sum_{k=1}^N \lambda_k y_k.$$

Это сразу следует из того, что зазор двойственности оценивает сверху разность между получившимися значениями в прямой задачи и двойственной [24]. Эта разность всегда неотрицательна, и на точных решениях прямой и двойственной задачи (и только на них) равна нулю. Заметим, что контроль онлайн-зазора двойственности

$$\Delta(\lambda; N) = \max_{u \in Q} \left\{ \frac{1}{S_k} \sum_{k=1}^N \lambda_k \langle \nabla f_k(x_k), x_k - u \rangle \right\}$$

позволяет в случае, когда удается выбрать $\lambda_k \equiv 1$, получать оценки регрета (псевдо регрета в стохастическом случае) в задачах онлайн оптимизации (см. [30, 40] и конец этого пункта). К сожалению, ограничение $\lambda_k \equiv 1$ существенно сужает класс методов. Скажем, для рассматриваемых в этом пункте быстрых градиентных методов $\lambda_k \sim k^p$. Кроме того, даже если в онлайн постановке допустить взвешивания с различными весами, все равно требуется, чтобы способ получения оценки на зазор двойственности допускал бы перенесения на онлайн-постановки. Быстрый градиентный метод, например, этого не допускает, что не сложно усмотреть из оценок работы [20].

Более общий способ оценки разности между получившимися значениями в прямой задачи и двойственной базируется на контроле сертификата точности [73, 74] (accuracy certificate), в который наряду с градиентами функционала входят градиенты нарушенных ограничений или в общем случае вектора нормалей к гиперплоскостям, отделяющим x_k

от множества Q (в ряде постановок “градиенты” стоит заменить на “субградиенты”). Собственно, такая интерпретация двойственных множителей следует из способа обоснования принципа множителей Лагранжа на основе следствия теоремы Хана–Банаха (теоремы об отделимости) [75]. Причем в работе [74] за счет слейтеровской релаксации ограничений (допущения возможности нарушения ограничений на ε [9, 62]) получаются оценки скорости сходимости, не зависящие от размера двойственного решения, который может быть большим.

Во многих (транспортно-)экономических приложениях при поиске равновесных конфигураций требуется решать пару задач (прямую и двойственную), см., например, [47, 69–72]. Если у этой пары задач, на которую можно смотреть, как на одну седловую задачу, есть определенная структура (проявляющаяся, например, в сепарабельности ограничений или в сильной выпуклости, см. также пример 4 и [62]), то упоминаемые ранее нижние оценки могут быть улучшены.

В действительности, выбранный в данной статье класс проекционных методов с построением модели функции далеко не единственный возможный способ строить прямо-двойственные методы. Скажем, уже упоминавшиеся методы условного градиента также являются прямо-двойственными [16, 17]. Еще более удивительным может показаться, что прямо-двойственная интерпретация есть, например, у метода эллипсоидов [73]. Более того, в ряде ситуаций мы можем за линейное время (с геометрической скоростью сходимости) находить одновременно решение прямой и двойственной задачи. Причем речь идет не только о конструкциях типа [76], базирующихся на принципе: сопряженная функция к выпуклой функции с липшицевым градиентом – сильно выпуклая, и обратно, сопряженная к сильно выпуклой функции – выпуклая функция с липшицевым градиентом.

В связи со сказанным выше интересно заметить, что если множество $Q \subset \mathbb{R}^n$ есть шар $B_q^n(R)$ радиуса R в q -норме,¹⁶ то нижние оценки (для случая $D = \delta = 0$) на точность (по функции), которую можно получить после $N \leq n$ итераций, имеют вид [77] (напомним, что¹⁷ $\|\nabla f(x) - \nabla f(y)\|_q \leq L_\nu \|x - y\|_{q'}^\nu$, $1/q + 1/q' = 1$, $\nu \in (0, 1]$):

¹⁶ Если о выпуклом замкнутом множестве Q известно только то, что оно содержит $B_q^n(R)$, то все сказанное далее также остается в силе.

¹⁷ В случае достаточной гладкости функции $f(x)$ можно выписать следующее представления для константы Липшица градиента (верхний индекс q соответствует выбору нормы в прямом пространстве)

$$L^q = \max_{x \in Q, \|h\|_q \leq 1} \langle h, \nabla^2 f(x)h \rangle.$$

$$\Omega\left(\frac{1}{\min\{q, \ln n\}^\nu} \frac{L_\nu R^{1+\nu}}{N^{\nu+(\nu+1)/q}}\right) (2 \leq q \leq \infty),$$

$$\Omega\left(\frac{1}{\ln^\nu(N+1)} \frac{L_\nu R^{1+\nu}}{N^{\nu+(\nu+1)/2}}\right) (1 \leq q < 2).$$

Приведенный результат хорошо соответствует тому, что написано в замечании 2.

Для $q = \infty$ и $\nu = 1$ (гладкий случай) приведенная оценка с точностью до логарифмического фактора будет иметь вид $\Omega(LR^2/N)$. Эта оценка достигается, например, на методе условного градиента Франка–Вульфа [16, 78].¹⁸ Исходя из только что написанного и тезиса о неулучшаемости оценок (7) ((8), (9)) (при $D = \delta = \mu = 0$, $p = 1$) может возникнуть ощущение противоречия. Это ощущение дополнительно усиливается примером 2 из п. 2. Действительно, исходя из этого примера, может сложиться ощущение, что проблема выбора прокс-структурь в задаче не очень актуальна, поскольку при проектировании можно исходить просто из самой нормы. И это действительно так, если мы ограничиваемся не ускоренными градиентными методами (PGM, метод Франка–Вульфа), которые сходятся как $O(L^q R_q^2 / N)$ (здесь $R_q = R$ – диаметр множества Q , посчитанный в q -норме, в нашем случае $q = \infty$). Если же мы хотим ускориться, и достичь оптимальной оценки $O(LR^2 / N^2)$, то уже необходимо существенно использовать проск-функцию $d(x) \geq 0$ со свойством сильной выпуклости относительно выбранной нормы и с константой сильной выпуклости $\alpha \geq 1$ [13]. Скажем (в связи с примером 2), квадрат 1-нормы – не есть сильно выпуклая функция относительно 1-нормы, т.е. $d(x) = \|x\|_1^2$ – не есть проск-функция с 1-нормой для симплекса. В классе удовлетворяющих условию 1-сильной выпуклости прокс-функций подбирается такая, которая минимизирует $R^2 := \max_{x \in Q} d(x)$. Именно это R^2 входит в оценку FGM $O(LR^2 / N^2)$. И как уже отмечалось (см. п. 2) для $Q = B_\infty^n(R)$ имеет место следующая оценка на прокс-диаметр $R^2 := \max_{x \in Q} d(x) = R^2 \Omega(n)$. Отсюда, с учетом того, что $N \leq n$, получаем, что оценка

В частности, $L^1 \leq L^2 \leq nL^1$, $L^2 \leq L^\infty \leq nL^2$. Эти формулы вместе со сказанным ранее относительно того, как может меняться M – обычная константа Липшица $f(x)$, при изменении нормы в прямом пространстве, поясняют почему в “устойчивые сочетания” эти константы входят таким образом: $M^2 R^2 / \varepsilon^2$, LR^2 / ε .

¹⁸ Отметим также, что этот метод допускает обобщение на случай неточного оракула, и неулучшаемость оценок может быть проинтерпретирована с точки зрения сохранения свойства разреженности решения [16].

$O(LR^2/N)$ и оценка $O(LR^2n/N^2)$, приводят, в общем-то, к одному результату, но в случае использования FGM требуется дополнительно искать оптимальную прокс-структуру. Только в таком случае будет совпадение результатов. Более того, также как и в замечании 2, здесь хорошо видно, что при $q \geq 2$ можно ограничиться рассмотрением только евклидовой прокс-структуры для FGM и евклидовой нормы для метода Франка–Вульфа. В частности, для $Q = B_\infty^n(R)$ действуя так, мы получим оценку $O(L^2R^2n/N^2)$ вместо ранее полученной оценки $O(L^\infty R^2n/N^2)$, соответствующей при $N \leq n$ неулучшаемой оценке $O(L^\infty R^2/N)$ (здесь мы проставили верхние индексы у L , поскольку они различаются). Дальше можно написать все тоже по поводу неулучшаемости оценок, что и в конце замечания 2.

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 и в ряде случаев условий остановки. Из этой ситуации можно выйти за логарифмическое (по этим параметрам) число рестартов метода. Стартуя, скажем, с $R=1$ и делая, предписанное этим R число шагов, мы проверяем выполняется ли для вектора, выдаваемого алгоритмом, условие ε -близости по функции (при условии, что мы можем сделать такую проверку). Если условие ε -близости не выполняется, то полагаем $R := 2R$ и т.д. Все эти перезапуски увеличат общее число обращений к оракулу лишь в $O(1)$ раз.¹⁹ Аналогичное можно сказать про²⁰ L , M и D . Однако если убрать стохастичность (и не рассчитывать на сильную выпуклость), тогда L , M можно не только эффективнее адаптивно подбирать (аналогично правилу Армихо [15] в независимости от того можем ли мы сделать проверку условия ε -близости значения функции в текущей точке к оптимальному) по ходу самих итераций (увеличив в среднем число итераций не более чем в 4 раза), но и в некотором смысле оптимально самонастраиваться (используя формулу (10)) на гладкость функционала на текущем участке пребывания метода [79]. Это означает, что в

¹⁹ В свою очередь можно поиграть и на этом $O(1)$, стараясь его минимизировать. Для этого шаг, который мы для простоты положили равным 2, подбирают оптимально исходя из того, с каким показателем степени входит прогоняемый параметр в оценку числа итераций [62].

²⁰ Впрочем, в детерминированных постановках мы можем явно наблюдать за последовательностью выдаваемых оракулом субградиентов и отслеживать условие на норму субградиентов. Как только наше предположение нарушилось (при этом мы не успели сделать предписанное текущему M число шагов), мы увеличиваем M в два раза и перезапускаем весь процесс с новым значением M . Число таких перезапусков будет не более чем логарифмическим от истинного значения M . Все эти рассуждения с небольшими оговорками (типа равномерной п.н. ограниченности стохастического субградиента) переносятся и на стохастические постановки, в которых наблюдается стохастический субградиент.

детерминированном случае без учета сильной выпуклости функционала существует универсальный метод, работающий по оценкам (8) с L , рассчитанной по формуле (10) (в которой δ берется из (9)), и оптимальным в смысле скорости сходимости выбором параметра $\nu \in [0,1]$. Причем выбор ν осуществляется не нами заранее, исходя из знания всех констант и минимизации выписанных оценок, а самим алгоритмом (здесь выбрано $p=1$):

$$N(\varepsilon) = \inf_{\nu \in [0,1]} \left(\frac{2^{\frac{3+5\nu}{2}} L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+3\nu}}.$$

Это соответствует (с точностью до логарифмического фактора) нижним оценкам [73], выписанным выше для случая $q \in [1,2]$. Отметим, что здесь при определении R используется соответствующая прокс-функция, см. замечание 2. К сожалению, пока не очень понятно ли что-то похожее сделать с параметром μ , с введенным нами параметром метода $p \in [0,1]$, и можно ли идеи универсальности (самонастраиваемости) распространить на задачи стохастической оптимизации. Обзор других работ на тему самонастройки алгоритмов в гладком детерминированном случае имеется в [19].

Приведенную оценку можно обобщить, если дополнительно известна μ -сильная выпуклость функции $f(x)$. Можно дополнительно к искусственно введенной игре на неточности оракула допустить, что имеет место и настоящая неточность. В этом случае также можно выписать соответствующие оценки. Не играя на выборе $\nu \in [0,1]$, можно распространить все, что описано выше в этом абзаце на стохастические постановки. Аналогичное можно сделать для стохастических безградиентных и покомпонентных методов с неточным оракулом (см. п. 4). Соответствующие обобщающие формулы собраны в работе [80], мы не будем их здесь приводить.

Выше мы сделали обременительную оговорку о возможности выполнять проверку условия ε -близости по функции. Такое заведомо возможно только при известном значении функционала в точке оптимума. Как правило, такой информации у нас априорно нет. Один из способов выхода из этой ситуации для задач стохастической оптимизации описан в п. 7.7 работы [53]. Отметим также, что в гладком детерминированном μ -сильно выпуклом случае, когда в точке минимума x_* выполняется условие $\nabla f(x_*) = 0$ (в частности, в задачах безусловной минимизации), критерий ε -близости может быть переписан в терминах малости (рассчитываемого) на итерациях градиента:

$$f(x_k) - f_* \leq \frac{1}{2\mu} \|\nabla f(x_k)\|_2^2 \leq \varepsilon.$$

В случае, когда имеется явная формула, выражающая решение прямой задачи через двойственные множители (сюда относятся, например, задачи энтропийно-линейного программирования), также можно оценивать точность решения прямой задачи (даваемой этими явными формулами) по точности решения двойственной задачи [62], используя неравенство²¹

$$\frac{1}{2L} \|\nabla f(x_k)\|_2^2 \leq f(x_k) - f_*.$$

Это обстоятельство используется в критерии остановки двойственного метода из [62].

Все сказанное выше, по-видимому, переносится в полной мере на задачи композитной оптимизации [13, 52, 81] и некоторые их обобщения, например, [63, 67, 68].

Также нам видится, что сказанное выше переносится на седловые задачи и монотонные вариационные неравенства [4, 13]. Причем, речь идет не о том, что было описано в примере 4, а о том, как скажется неточность оракула на оптимальные методы для седловых задач и монотонных вариационных неравенств [4, 13]. Ответ, более менее, известен: неточность оракула не будет накапливаться для оптимальных методов (в отличие от задач обычной выпуклой оптимизации). Отметим, что концепцию неточного оракула еще необходимо должным образом определить²² – предположение 1 нуждается в корректировке для данного класса задач. Отсутствие накопления неточностей связано с тем, что для таких задач оценка (7) будет оптимальна (с некоторыми оговорками) при $p=0$. Другие $p \in (0,1]$ рассматривать не стоит (такие оценки просто не достижимы). Впрочем, пока про это имеются лишь частичные результаты [63, 67, 68].

В число итераций описанных методов не входит явно размерность пространства n . Это наталкивает на мысль (подобные мысли, по-видимому, впервые были высказаны и реализованы для класса обычных градиентных методов в кандидатской диссертации Б.Т. Поляка [1], см также [14]) о возможности использовать эти методы, например, в гильбертовых пространствах [15, 39]. Оказывается это, действительно, можно делать при

²¹ В замечании 5 был приведен пример, когда $\|\nabla f(x_k)\|_2 = O(k^{-2})$. Из выписанного неравенства мы можем гарантировать лишь $\|\nabla f(x_k)\|_2^2 = O(k^{-2})$. Ситуацию можно улучшить, если регуляризовать функционал (см. конец п. 2 и [62]), сделав его сильно выпуклым, и применить FGM [10, 13] к регуляризованной задаче, тогда $\|\nabla f(x_k)\|_2 = O((\ln k/k)^2)$. Нам не известно, можно ли в общем случае избавится от этого логарифмического фактора. В негладком случае ситуация проще, см. [24].

²² Например, для монотонного вариационного неравенства: найти такой $x \in Q$, что для всех $y \in Q$ выполняется $\langle g(y), y - x \rangle \geq 0$, достаточные условия на (δ, L) -оракул будут иметь вид: для любых $x, y \in Q$

$$\langle g(y) - g(x), y - x \rangle \geq -\delta, \|g(y) - g(x)\|_* \leq L\|x - y\| + \delta.$$

Вероятно, в ряде ситуаций эти условия можно ослабить (доказательства в этом случае нам не известны)

$$-\delta \leq \langle g(y) - g(x), y - x \rangle \leq L\|y - x\|^2 + \delta.$$

определенных условиях (см., например, [80] в контексте использованных в данной работе обозначений). В частности, концепция неточного оракула позволяет привнести сюда элемент новизны, существенно мотивированный практическими нуждами – принципиальной невозможностью (в типичных случаях нет явных формул) решать с абсолютной (очень хорошей) точностью вспомогательную задачу на каждом шаге градиентного спуска. Например, решение такой вспомогательной задачи для класса задач оптимального управления приводит к краевой задаче, которую необходимо решить для вычисления градиента функционала. Однако, в действительности, почти все практически интересные задачи (за редким исключением, к коим можно отнести класс Ляпуновских задач [75]) в бесконечномерных пространствах не являются выпуклыми, поэтому здесь имеет смысл говорить лишь о локальной теории [82]. Отметим также, что концепция неточного оракула намного плодотворнее может использоваться в данном контексте, если аппроксимировать бесконечномерную задачу конечномерной, в которой размерность конечномерной задачи стоит выбирать исходя из игры на качестве аппроксимации и на том, как входит в оценки общих вычислительных затрат эта размерность: чем больше размерность, тем лучше аппроксимация, но тем больше требуется итераций (впрочем, здесь зависимость может быть слабая) и тем дороже одна итерация. Выписанные в данном пункте оценки играют ключевую роль в этой игре.

Покажем в заключении этого пункта, как приведенные результаты переносятся на задачи стохастической онлайн оптимизации. Для этого напомним вкратце, в чем состоит постановка задачи (см., например, [23, 30, 40, 83–88]). Требуется подобрать последовательность $\{x^k\} \in Q$ так, чтобы минимизировать псевдо регрет:

$$\frac{1}{N} \sum_{k=1}^N E_{\xi^k} \left[f_k(x^k, \xi^k) \right] - \min_{x \in Q} \frac{1}{N} \sum_{k=1}^N E_{\xi^k} \left[f_k(x, \xi^k) \right]$$

на основе доступной информации $\{\nabla \tilde{f}_1(x^1, \xi^1); \dots; \nabla \tilde{f}_{k-1}(x^{k-1}, \xi^{k-1})\}$ при расчете x^k , где

$$\|\nabla \tilde{f}_i(x^i, \xi^i) - \nabla f_i(x^i, \xi^i)\|_* \leq \delta.$$

Здесь с.в. $\{\xi^k\}$ могут считаться независимыми одинаково распределенными. Онлайнность постановки задачи допускает, что на каждом шаге k функция f_k может подбираться из рассматриваемого класса функций враждебно по отношению к используемому нами методу генерации последовательности $\{x^k\}$. В частности, f_k может зависеть от

$$\{x^1, \xi^1, f_1(\cdot); \dots; x^{k-1}, \xi^{k-1}, f_{k-1}(\cdot); x^k\}.$$

В стохастической онлайн оптимизации с неточным оракулом можно получить следующие оценки псевдо регрета (см., например, [86], случай неточного оракула в похожем контексте ранее уже частично прорабатывался в [28])

$$(11) \quad \min \left\{ O\left(\sqrt{\frac{M^2 R^2}{N}} + \delta\right), O\left(\frac{M^2 \ln N}{\mu N} + \delta\right) \right\},$$

где $\|\nabla f_k(x, \xi)\|_* \leq M$ – равномерно по x и п.н. по ξ . Эти оценки достигаются (фактически на тех же методах, что и в п. 2 с небольшой оговоркой в сильной выпуклом случае [36, 86]) и неулучшаемы (в том числе для детерминированных постановок с $\delta = 0$). Как видно из этих оценок, наличие гладкости ничего не дает. То есть никакого аналога формулы (7) здесь нет. Все что ранее говорилось про прокс-структуру и большие отклонения, насколько нам известно, полностью и практически без изменений (в μ -сильно выпуклом случае в оценках вероятностей больших уклонений $\ln(\ln(N)) \rightarrow \ln N$ – доказательство это утверждения мы не смогли найти) переносится и на онлайн случай.

4. Стохастические безградиентные и покомпонентные методы с неточным оракулом

Рассматривается задача стохастической выпуклой оптимизации (1)

$$f(x) = E_\xi[f(x, \xi)] \rightarrow \min_{x \in Q} .$$

Предположения те же, что и в первом абзаце п. 2. В частности, п.н. $\|\nabla f(x, \xi)\|_2 \leq M$. Здесь важно, что функция $f(x)$ задана не только на множестве Q , но и в его τ_0 -окрестности (см. ниже), и все предположения делаются не для $x \in Q$, а для x из τ_0 -окрестности множества Q (аналогичная оговорка потребуется далее, при перенесении результатов п. 3). Однако теперь оракул не может выдавать стохастический субградиент функции. На каждой итерации мы можем запрашивать у оракула только значения реализации функции $f(x, \xi)$ в нескольких точках x . Принципиальная разница есть только между одной или двумя точками [80, 89]. Здесь мы ограничимся рассмотрением случая двух точек – случай одной точки представляет интерес только в онлайн контексте (см. [88] и цитированную там литературу). Впрочем, есть достаточно большой и популярный класс одноточечных не онлайн постановок, который мы здесь не касаемся (см., например, [90, 91]).

Предположение 2. δ -оракул выдает (на запрос, в котором указывается только одна точка x) $f(x, \xi) + \delta(x, \xi)$, где с.в. ξ независимо разыгрывается из одного и того же распределения, фигурирующего в постановке (1), случайная величина

$\delta(x, \xi) = \tilde{\delta}(x) + \bar{\delta}(\xi)$, где $\bar{\delta}(\xi)$ – независимая от x случайная величина с неизвестным распределением (случайность которой может быть обусловлена не только зависимостью от ξ), ограниченная по модулю $\delta/2$ (число δ – допустимый уровень шума), $\tilde{\delta}(x)/(R\delta)$ – неизвестная 1-липшицева функция.

Далее в изложении мы будем во многом следовать [2, 31, 40, 50, 86–92]. По полученным от оракула зашумленным значениям мы можем определить стохастический субградиент:

$$(12) \quad g_{\tau, \delta}(x, s, \xi) = \frac{n}{\tau} (f(x + \tau s, \xi) + \delta(x + \tau s, \xi) - (f(x, \xi) + \delta(x, \xi))) s,$$

где s – случайный вектор (независимый от ξ), равномерно распределенный на $S_2^n(1)$ – единичной сфере в 2-норме в пространстве \mathbb{R}^n .²³ Из этого представления можно усмотреть, что липшицева составляющая шума δ_2 из предположения 2 и уровень шума δ_1 из предположения 1 связаны соотношением $\delta_2 \sim \delta_1/n$ (см. формулу (24)). В действительности, для обоснования этой связи требуются значительно более громоздкие рассуждения.

Приведем одну из возможных мотивировок введенной в предположении 2 концепции δ -оракула. Предположим, что оракул может считать абсолютно точно значение липшицевой функции, но вынужден нам выдавать лишь конечное (предписанное) число первых бит. Таким образом, в последнем полученном бите есть некоторая неточность (причем мы не знаем по какому правилу оракул формирует этот последний выдаваемый значащий бит). Однако мы всегда можем прибавить (по mod 1) к этому биту случайно приготовленный (независимый) бит. В результате, не ограничивая общности, можно считать, что оракул последний бит выбирает просто случайно в независимости от отброшенного остатка. То что в итоге выдает оракул соответствует концепции δ -оракула.

Перейдем к получению оценок. В отличие от пп. 2, 3 везде далее в этом пункте мы будем считать, что имеет место обратное неравенство на требуемое число итераций $N \geq n$ [89]. Прежде всего, заметим, что²⁴

²³ С помощью леммы Пуанкаре такой вектор можно сгенерировать за $O(n)$, разыгрывая n независимых одинаково распределенных стандартных нормальных случайных величин и нормируя их на корень из суммы их квадратов.

²⁴ Взятие математического ожидания по δ подчеркивает, что $\delta(x, \xi)$ может быть случайной величиной не только потому, что может зависеть от ξ , но и потому, что может иметь собственную случайность.

$$(13) \quad E_{s,\xi,\delta} [g_{\tau,\delta}(x, s, \xi)] = \nabla f_\tau(x) + \nabla_x E_{\tilde{s},\xi,\delta} [\delta(x + \tau \tilde{s}, \xi)],$$

где \tilde{s} – случайный вектор, равномерно распределенный на $B_2^n(1)$ – единичном шаре в 2-норме, а $f_\tau(x) = E_{\tilde{s},\xi} [f(x + \tau \tilde{s}, \xi)]$ – сглаженная²⁵ версия функции $f(x) = E_\xi [f(x, \xi)]$. Причем,

$$(14) \quad 0 \leq f_\tau(x) - f(x) \leq M\tau,$$

$$(15) \quad \|g_{\tau,\delta}(x, s, \xi)\|_2 \leq n \left(M + \frac{2\delta}{\tau} \right).$$

Основная идея [2] заключается в подмене задачи (1) следующей задачей

$$(16) \quad f_\tau(x) = E_{\tilde{s},\xi} [f(x + \tau \tilde{s}, \xi)] \rightarrow \min_{x \in Q},$$

$\varepsilon/2$ -решение которой при $\tau = \varepsilon/(2M)$, будет ε -решением исходной задачи (1).

Считая $\delta = O(\varepsilon)$ и²⁶ (приведенное условие выполняется, если мы имеем доступ к δ -оракулу из предположения 2)

$$\|\nabla_x E_{\tilde{s},\xi,\delta} [\delta(x + \tau \tilde{s}, \xi)]\|_2 = O(\varepsilon/R),$$

можно получить для среднего числа итераций (используя те же алгоритмы для задачи (16), что и в п. 2, со стохастическим градиентом (12)), соответствующие аналоги оценок (2), (4):

$$(17) \quad O(n^2 M^2 R^2 / \varepsilon^2), O(n^2 M^2 / (\mu \varepsilon)).^{27}$$

²⁵ Все свойства функции $f(x)$ при переходе к $f_\tau(x)$ могут только улучшиться. В частности, $f_\tau(x)$ также выпуклая функция (можно перенести и на сильную выпуклость с не меньшей константой), с константой Липшица и константой Липшица градиента (если таковая существует у $f(x)$) не большей чем у $f(x)$.

²⁶ Если это условие не выполняется, то все что написано далее останется верным, правда, при более ограничительных условиях на допустимый уровень шума (это касается и всего последующего изложения). Так, если не налагать это ограничение, то потребуется считать $\delta = O(\varepsilon^2 / (\sqrt{n}MR))$ или $\delta = O(\varepsilon^{3/2} / (\sqrt{n}LR))$ – в случае, если $f(x)$ имеет L -липшицев градиент. Это можно получить с помощью замечания 6.

²⁷ Аналогично (2), (4) можно переписать оценку (17) не в среднем, как сейчас, а с учетом вероятностей больших уклонений. Это замечания касается и последующих вариаций формулы (17). Нам не известно

Если дополнительно известно, что $f(x)$ – гладкая функция и

$$(18) \quad \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2,$$

то вместо (14) будем иметь

$$(19) \quad 0 \leq f_\tau(x) - f(x) \leq \frac{L\tau^2}{2}.$$

Из формулы (19) следует, что можно ослабить требование к неточности: допускать неточность оракула масштаба²⁸ $\delta \sim M\tau \sim M\sqrt{\varepsilon/L}$.

При сделанных дополнительных условиях гладкости (18) за счет ужесточения требований к масштабу допускаемой неточности δ (как именно требуется это сделать можно усмотреть из формулы (21); ниже мы вернемся к этому вопросу) можно улучшить скорость сходимости (фактор n^2 перейдет в n):

$$(20) \quad O(nM^2R^2/\varepsilon^2), O(nM^2/(\mu\varepsilon)).$$

Оценки (20) в общем случае не улучшаются (даже при $\delta = 0$) для гладких стохастических и негладких задач [89]. Фактически это означает, что мы можем выбрать настолько малое τ (насколько малым мы можем его выбрать определяется ε и δ), что конечная разность в (12) “превращается” (с нужной точностью) в производную по направлению. Для объяснения отмеченного перехода полезно заметить, что [50, 87–89] (см. также (22))

$$(21) \quad E_{s,\xi} \left[\|g_{\tau,\delta}(x, s, \xi)\|_2^2 \right] \leq 4nM^2 + L^2\tau^2n^2 + \frac{8\delta^2n^2}{\tau^2}.$$

Замечание 6 (техника двойного сглаживания негладких задач Б.Т. Поляка [2], см. также [89]). За счет подмены изначально негладкого функционала в задаче (1) на

$$f(x) := f_\gamma(x) = E_{\tilde{s}_1, \eta} [f(x + \gamma \tilde{s}_1, \xi)], \gamma = \varepsilon/(2M),$$

являются ли оценки (17) оптимальными при заданном уровне шума $\delta = O(\varepsilon)$. Далее будет показано (замечание 6), что если налагать более жесткие условия на уровень шума, то оценки (17) можно улучшить.

²⁸ Здесь мы дополнительно считаем, что $\nabla_x E_{\tilde{s}, \xi, \delta} [\delta(x + \tau \tilde{s}, \xi)] = 0$. В частности, это условие выполняется, если неточность $\delta(x, \xi)$ имеет независимое от x распределение.

где \tilde{s}_1 – случайный вектор (независимый от ξ), равномерно распределенный на $B_2^n(1)$, получим новую задачу ($\varepsilon/2$ -аппроксимирующую исходную), которая уже будет гладкой с $L = O(nM^2/\varepsilon)$, см. [50]. Далее решая с помощью уже описанной техники с точностью $\varepsilon := \varepsilon/2$ задачу стохастической оптимизации (1) с

$$\xi := (\tilde{s}_1, \xi), f(x, \xi) := f(x + \gamma \tilde{s}_1, \xi),$$

$$g_{\tau, \delta}(x; s_2, \xi) := \frac{n}{\tau} (f(x + \gamma \tilde{s}_1 + \tau s_2, \xi) + \delta(x + \gamma \tilde{s}_1 + \tau s_2, \xi) - (f(x + \gamma \tilde{s}_1, \xi) + \delta(x + \gamma \tilde{s}_1, \xi))) s_2,$$

где s_2 – случайный вектор (независимый от ξ и \tilde{s}_1), равномерно распределенный на $S_2^n(1)$, получим те же оценки (20), только при еще более жестких условиях на уровень шума δ .

Оценки (17) и (20) переносятся и на задачи стохастической онлайн оптимизации (см., например, [86–89, 92]) с возникновением дополнительного фактора $\ln N$ в сильно выпуклом случае (см. формулу (11)). При этом в гладком случае не обязательно требовать дополнительно стохастичность исходной постановки для оптимальности оценок (20).

Далее рассматривается не стохастический вариант постановки задачи (1) (не обобщаемый на онлайн постановки) с $Q = \mathbb{R}^n$ (обобщения на произвольные выпуклые множества $Q \subset \mathbb{R}^n$ представляются интересными, но на данный момент нам не известны такие обобщения – в последующих рассуждениях существенным образом используется то, что в точке минимума $\nabla f(x) = 0$). Так что теперь R – расстояние от точки старта до решения (рассматривается коэрцитивный случай – см. замечание 4). В этом варианте, выписанная оценка (21) может быть уточнена

$$(22) \quad E_s \left[\|g_{\tau, \delta}(x, s)\|_2^2 \right] \leq 4n \|\nabla f(x)\|_2^2 + L^2 \tau^2 n^2 + \frac{8\delta^2 n^2}{\tau^2}.$$

Последняя оценка следует из явления концентрации равномерной меры на $S_2^n(1)$ с выделенными полюсами вокруг экватора (см. [93]):

$$E_s \left[\langle \nabla f(x), s \rangle^2 \right] = \frac{1}{n} \|\nabla f(x)\|_2^2.$$

Считая для простоты формулировок, что

$$\nabla_x E_{\tilde{s}, \delta} [\delta(x + \tau \tilde{s})] = 0,$$

можно распространить метод [60], дающий оценки (8), на текущий контекст, и получить следующие оценки (с $\tau \sim \sqrt{\delta/L}$)

$$(23) \quad N_1(\varepsilon) = n \cdot O\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p+1}}, \quad N_2(\varepsilon) = n \cdot O\left(\left(\frac{L}{\mu}\right)^{\frac{1}{p+1}} \ln\left(\frac{LR^2}{\varepsilon}\right)\right)$$

при (условия на допустимый уровень шума, при котором оценки (23) имеют такой же вид, с точностью до $O(1)$, как если бы шума не было²⁹)

$$(24) \quad \delta_1(\varepsilon) \leq \frac{1}{n} O\left(\varepsilon \cdot \left(\frac{\varepsilon}{LR^2}\right)^{\frac{p}{p+1}}\right), \quad \delta_2(\varepsilon) \leq \frac{1}{n} O\left(\varepsilon \cdot \left(\frac{\mu}{L}\right)^{\frac{p}{p+1}}\right).$$

По-видимому (строгим доказательством мы не располагаем на данный момент), и в стохастическом случае имеет место аналог формул (23), (24) с заменой в формуле (23)

$$N_1(\varepsilon) = n \cdot \max \left\{ O\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p+1}}, O\left(\frac{DR^2}{\varepsilon^2}\right) \right\}, \quad N_2(\varepsilon) = n \cdot \max \left\{ O\left(\left(\frac{L}{\mu}\right)^{\frac{1}{p+1}} \ln\left(\frac{LR^2}{\varepsilon}\right)\right), O\left(\frac{D}{\mu\varepsilon}\right) \right\}.$$

Можно продолжать переносить все написанное в п. 3 на рассматриваемую ситуацию (частично это уже сделано в [80]). Однако мы остановимся лишь на наиболее интересном (на наш взгляд) месте. А именно, на согласовании прокс-структурь с рандомизацией, порождающей сглаживание.

Основным результатом здесь является следующее наблюдение [88]: независимо от выбора прокс-структурь рандомизацию всегда стоит выбирать согласно (12). То есть с помощью разыгрывания случайного вектора s равномерно распределенного на $S_2^n(1)$, аналогичное относится и к замечанию 6. Ограничимся далее перенесением оценок (20) на общие прокс-структурь.

Приведем соответствующее обобщение формулы (21) (здесь и далее нижний индекс “2” у констант Липшица подчеркивает, что они считаются согласно евклидовой норме из-за сделанного нами выбора способа рандомизации)

²⁹ В отсутствии шума, оракул нам фактически может выдавать производную по направлению

$$g(x, s) = n \langle \nabla f(x), s \rangle s,$$

точнее $\langle \nabla f(x), s \rangle$, s мы генерируем сами. Если, в свою очередь, считать, что $\langle \nabla f(x), s \rangle$ оракул выдает с шумом (для простоты считаем, независящим от s) масштаба $\delta := \sqrt{L\delta}$ (δ в правой части определяется исходя из формулы (24)), то формула (23) останется верной.

$$E_{s,\xi} \left[\|g_{\tau,\delta}(x, s, \xi)\|_{q'}^2 \right] = O \left(\left(4M_2^2 n + L_2^2 \mu^2 n^2 + \frac{8\delta^2 n^2}{\mu^2} \right) E_s \left[\|s\|_{q'}^2 \right] \right),$$

где в прямом пространстве выбрана q -норма и $1/q + 1/q' = 1$. Согласно замечанию 2 можно считать, что $2 \leq q' \leq O(\log n)$ – выбирать другие нормы, как правило, бывает не выгодно.

Для такого диапазона $E_s \left[\|s\|_{q'}^2 \right] = \tilde{O}(n^{2/q'-1})$, в частности $E_s \left[\|s\|_{O(\log n)}^2 \right] = \tilde{O}(n^{-1})$ ($\tilde{O}(\cdot)$ с точностью до логарифмического фактора совпадает с $O(\cdot)$, аналогично с $\tilde{\Omega}(\cdot)$). Исходя из такого обобщения, можно привести следующую таблицу, распространяющую оценку (20) на произвольные прокс-структуры.

R – прокс-диаметр множества Q , с нормой $\ \cdot\ _q$ в прямом пространстве	$f(x)$ – выпуклая	$f(x)$ – μ_2 -сильно выпуклая
$E_\xi \left[f(x, \xi) \right] \leq B$	$O \left(\frac{nB^4 R^2}{\varepsilon^4 \tau_0^2} \right) \tilde{O}(n^{2/q-1})$	$O \left(\frac{nB^4}{\mu_2 \varepsilon^3 \tau_0^2} \right)$
$E_\xi \left[\ \nabla f_k(x, \xi)\ _2^2 \right] \leq M_2^2$	$O \left(\frac{nM_2^2 R^2}{\varepsilon^2} \right) \tilde{O}(n^{2/q-1})$	$O \left(\frac{nM_2^2}{\mu_2 \varepsilon} \right)$

Выпишем условия, из которых можно получить требования на шум (нам представляется, что здесь это может хорошо прояснить суть дела):

- $\min \{M_2 \tau, L_2 \tau^2 / 2\} = O(\varepsilon)$ – условие аппроксимации;
- $L_2^2 \tau^2 n^2 + \frac{8\delta^2 n^2}{\tau^2} = O(M_2^2 n)$ – условие “правильной” ограниченности квадрата нормы аппроксимации стохастического градиента;
- $M_2 = O \left(\frac{B^2}{\varepsilon \tau_0} \right)$ – условие на “эффективную” константу M_2 для первой строки таблицы [94], если имеющаяся константа Липшица $f(x)$ не известна или известна, но слишком большая;

- $L_2 = O\left(\frac{nM_2^2}{\varepsilon}\right)$ – условие на “эффективную” константу L_2 для второй строки таблицы (см. замечание 6), если $f(x)$ не имеет Липшицева градиента или имеющаяся константа Липшица градиента слишком большая.

Выписанные условия позволяют для всех полей таблицы (с оценками) написать соответствующие условия на допустимый уровень шума, и параллельно подобрать оптимальный размер шага τ .

Приведенные во втором столбце таблицы оценки при определенных условиях могут быть лучше нижних оценок [89]. Здесь ситуация аналогична той, о которой написано в конце замечания 2 (будем использовать те же обозначения) и в п. 3. Например, если $Q = B_1^n(1)$, то в нижней оценке [89] стоит $\tilde{\Omega}(nM_1^2/\varepsilon^2)(E_\xi[\|\nabla f(x, \xi)\|_\infty^2] \leq M_1^2)$, а в таблице будет стоять $\tilde{O}(M_2^2/\varepsilon^2)$. Осталось заметить, что $M_2^2 \leq nM_1^2$, причем в определенных ситуациях может быть $M_2^2 \ll nM_1^2$.

Основные конкурирующие рандомизации – это рандомизация на евклидовой сфере и покомпонентная рандомизация [88, 89]. Последняя осуществляется по принципу: равновероятно (за $O(\ln n)$) осуществляется выбор одной из координатных осей, затем независимо (за $O(1)$) осуществляется выбор одного из двух противоположных единичных ортов на этой оси. Покомпонентную рандомизацию можно организовать еще более эффективно с точки зрения последующих оценок скорости сходимости, если известна информация о свойствах функционала отдельно по каждой компоненте. Например, это можно сделать, когда известны константы Липшица компонент градиента функционала – рассматривается гладкий случай [43]. Кроме того, покомпонентные спуски хорошо параллелируются [48]. При таких дополнительных условиях описанная в этом пункте рандомизация на евклидовой сфере может уступить покомпонентной.

Приведем несколько примеров, демонстрирующих важность изучения безградиентных методов (часто эти методы называют прямыми методами [2] или методами 0-порядка [50]).

Пример 5 (двухуровневая оптимизация [95]). Требуется решить задачу, возникающую, например, при поиске равновесия по Штакельбергу [71]

$$\psi(x, u) \rightarrow \max_{u \in U},$$

$$f(x, u(x)) \rightarrow \min_x.$$

Из первой задачи находится зависимость $u(x)$, которая входит во вторую (внешнюю) задачу. Проблема здесь в том, что явная зависимость $u(x)$ в общем случае может быть недоступна. Как следствие, могут быть проблемы с расчетом $\nabla u(x)$. Поэтому предлагается приближенно решать первую задачу и использовать безградиентный метод с неточным оракулом для второй. Насколько точно надо решать первую задачу, и какой именно безградиентный метод (с точки зрения чувствительности к неточности) выбирать для второй – определяется сложностью решения первой задачи и свойствами второй.

Рассмотренная двухуровневая задача может быть сильно упрощена, если удается найти ее седловое представление [13, 67, 68]. В частности, если функции $\psi(x, u)$, $f(x, u)$ – выпуклы по x и вогнуты по u , и $\psi(x, u)$ – простой структуры, то можно (для достаточно большого λ) заменить исходную задачу на следующую

$$\min_x \max_{u \in U} [f(x, u) + \lambda \psi(x, u)].$$

Полученную седловую задачу стоит решать методами композитной оптимизации [13, 67, 68], чтобы λ либо совсем не входила в оценки числа итераций, либо входило очень слабо.

К сожалению, седловое представление возможно далеко не всегда.

Пример 6 (огромная скрытая размерность). Пусть $y(x) \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $n \ll m$. Требуется решить задачу

$$f(x, y(x)) \rightarrow \min_x .$$

Мы считаем, что мы можем эффективно посчитать с необходимой точностью $y(x)$ и $f(x, y(x))$ за $O(m)$. Если для решения этой задачи оптимизации мы будем использовать безградиентный метод, то общее число сделанных арифметических операций пропорционально mn (см. (20), (23)), но если мы будем использовать обычный градиентный метод, то общее число сделанных арифметических операций пропорционально m^2n .

В действительности, часто имеет место следующее полезное наблюдение [96]: если мы можем вычислить значения $y(x)$ и $f(x, y(x))$ за $O(m)$, то мы можем с такой же по порядку сложностью вычислить и производные по фиксированному направлению h :

$$\frac{dy(x)}{dx} h, \left\langle \frac{\partial f(x, y)}{\partial x}, h \right\rangle + \left\langle \frac{\partial f(x, y)}{\partial y}, \frac{dy(x)}{dx} h \right\rangle.$$

Тем не менее, тут требуется много оговорок, в том числе про точность расчетов. Если не вдаваться в детали, то такие рассуждения также приводят к затратам пропорциональным tn , где n возникло в виду оценок (20), (23) для спусков по направлению. Оговорки о точности здесь все же необходимы, поскольку для безградиентных методов и спусков по направлению требования к точности могут существенно отличаться (об этом ранее уже было немного написано в данном пункте). Как следствие, в оценку $O(m)$ необходимо явно вводить зависимость от точности вычисления $u(x)$ и $f(x, u(x))$. Об этом планируется написать отдельно.

В примерах 5, 6, в действительности, требуются некоторые оговорки о невозможности или неэффективности использования БАД (см. п. 2, а также [38, 39]). Мы знаем несколько вполне конкретных случаев (один из них возник из примеров 1, 2 п. 2, см. [96]), в которых БАД полностью нивелирует те сложности, которые были описаны в этих примерах. Но нам также известны случаи, подпадающие под разобранные примеры, в которых пока не очень понятно, как можно было бы воспользоваться БАД [71]. Кроме того, в этих примерах важно уметь эффективно пересчитывать значения $u(x)$ или $y(x)$, а не рассчитывать их каждый раз заново (на каждой итерации внешнего цикла). Поясним сказанное. Предположим, что мы уже как-то посчитали, скажем, $u(x)$, решив с какой-то точностью соответствующую задачу оптимизации. Тогда для вычисления $u(x + \Delta x)$ (на следующей итерации внешнего цикла) у нас будет хорошее начальное приближение $u(x)$. А как известно (см. пп. 2–4) расстояние от точки старта до решения (не в сильно выпуклом случае) существенным образом определяет время работы алгоритма. Тем не менее, известные нам приложения (пример 4 п. 3 и [96]) пока как раз всецело соответствуют сильно выпуклой ситуации. Связано это с тем, что если расчет $u(x)$ или $y(x)$ с точностью ε осуществляется за $O(C \ln(R/\varepsilon))$ операций, то для внешней задачи можно выбирать самый быстрый метод (а, стало быть, и самый требовательный к точности), и с точностью до того, что стоит под логарифмом, общая трудоемкость будет просто прямым произведением трудоемкостей решения внутренней и внешних задач по отдельности. Как правило, такое сочетание оказывается недоминируемым.

Также необходимо отметить, что, как правило, итоговые задачи оптимизации (после подстановки зависимости $u(x)$ или $y(x)$) в этих примерах получаются не выпуклыми. В этой связи можно лишь говорить о локальной теории. В отсутствии выпуклости даже если ограничиться локальной сходимостью многое из того, что описано в данной статье, требует отдельного рассмотрения [50, 63].

Отметим в заключении, что если немного по-другому посмотреть на описанное в этом пункте, то можно заметить следующее. Какой бы большой (но равномерно

ограниченный по итерациям) шум ни был, если $\delta(x + \tau s, \xi)$ имеет распределение не зависящее от s , то (возможно через очень большое число итераций) мы сможем сколь угодно точно (по функции) решить задачу! Аналогичное можно сказать, если мы изначально исходим из концепции оракула, выдающего зашумленное значение $\langle \nabla f(x), s \rangle$, причем зашумленность не зависит от s . Все это восходит к идеям Р. Фишера, развитым О.Н. Границиним и Б.Т. Поляком [31].

Авторы выражают благодарность А.Ю. Горнову, Ф. Глинёру, О.Н. Границину, О. Деволдеру, Ю.Г. Евтушенко, Н.К. Животовскому, А.В. Назину, А. Рахлину, П. Рихтарику, А. Содину, В.Г. Спокойному, С.П. Тарасову, Э. Хазану, А.Б. Юдицкому и особенно А.С. Немировскому за возможность обсуждения разных частей данного текста в разное время, а также Ильнуре Усмановой, обнаружившей несколько опечаток.

Особо хотелось бы отметить то позитивное влияние, которое на каждого из авторов оказал Борис Теодорович Поляк, отмечающий в этом году свое восьмидесятилетие. Так Ю.Е. Нестеров защищал кандидатскую диссертацию под руководством Б.Т. Поляка в 1983 г. Два других соавтора в значительной степени являются учениками Ю.Е. Нестерова по части, связанной с численными методами выпуклой оптимизации. Набор идей, здоровый оптимизм, понятность и открытость Бориса Теодоровича всегда делает общение (работу) с ним чрезвычайно интересной и полезной. Настоящая статья во многом была вдохновлена его книгой [2], и попыткой ее переосмыслить с современных позиций. Несмотря на то, что большей частью эта книга была написана почти 40 лет назад, она до сих пор не утратила свою актуальность, о чем говорит, например, выход ее нового издания в прошлом году.

Работа выполнена при поддержке грантов РФФИ 13-01-12007-офи_м, 14-01-00722-а, Лаборатории структурных методов анализа данных в предсказательном моделировании ФУПМ МФТИ, грант правительства РФ дог. 11.G34.31.0073. Исследования, проведенные в п. 4, были выполнены за счет гранта Российского научного фонда (проект № 14-50-00150).

СПИСОК ЛИТЕРАТУРЫ

1. Поляк Б.Т. Градиентные методы минимизации функционалов, решения уравнений и неравенств. Диссертация на соискание ученой степени кандидата физико-математических наук. МГУ, механико-математический факультет, 1963.
2. Поляк Б.Т. Введение в оптимизацию. М.: Наука, 1983.
3. Хачиян Л.Г. Избранные труды / сост. С. П. Тарасов. М.: МЦНМО, 2009.
4. Немировский А.С., Юдин Д.Б. Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.
http://www2.isye.gatech.edu/~nemirovs/Lect_EMCO.pdf
5. Поляк Б.Т., Цыпкин Я.З. Оптимальные псевдоградиентные алгоритмы адаптации // Автоматика и телемеханика. 1980. № 8. С. 74–84.
6. Цыпкин Я.З., Позняк А.С. Оптимальные поисковые алгоритмы стохастической оптимизации // Докл. АН СССР. 1981. Т. 260. № 3. С. 550–553.
7. Нестеров Ю.Е. Эффективные методы в нелинейном программировании. М.: Радио и связь, 1989.
8. Nesterov Y., Nemirovsky A. Interior-point polynomial methods in convex programming. Studies in applied mathematics. V. 13. SIAM, Philadelphia, 1994.
9. Nemirovski A. Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2013.
http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf
10. Нестеров Ю.Е. Введение в выпуклую оптимизацию. М.: МЦНМО, 2010.
11. Bubeck S. Theory of convex optimization for machine learning // e-print, 2014. [arXiv:1405.4980](https://arxiv.org/abs/1405.4980)
12. <http://cvxr.com/cvx/>
13. Нестеров Ю.Е. Алгоритмическая выпуклая оптимизация. Диссертация на соискание степени д.ф.-м.н. по специальности 01.01.07 – вычислительная математика. Долгопрудный, МФТИ 26 декабря 2013 г. 367 стр.
14. Левитин Е.С., Поляк Б.Т. Методы минимизации при наличии ограничений // ЖВМ и МФ. 1966. Т. 6. № 5. С. 787–823.
15. Васильев Ф.П. Методы оптимизации. М.: МЦНМО, 2011.
16. Jaggi M. Revisiting Frank–Wolfe: Projection-free sparse convex optimization // Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013.
17. Nesterov Yu. Complexity bounds for primal-dual methods minimizing the model of objective function // CORE Discussion Papers. 2015/03. 2015.
18. Tseng P. On accelerated proximal gradient methods for convex-concave optimization // SIAM J. Optim. 2008. (submitted) <http://www.mit.edu/~dimitrib/PTseng/papers.html>
19. Neumaier A. OSGA: A fast subgradient algorithm with optimal complexity // e-print, 2014. [arXiv:1402.1125](https://arxiv.org/abs/1402.1125)
20. Allen-Zhu Z., Orecchia L. Linear coupling: An ultimate unification of gradient and mirror descent // e-print, 2014. [arXiv:1407.1537](https://arxiv.org/abs/1407.1537)
21. Ермольев Ю.М. Методы стохастического программирования. М.: Наука, 1976.
22. Nesterov Y. Minimizing functions with bounded variation of subgradients // CORE Discussion Papers. 2005/79. 2005.

23. Гасников А.В., Нестеров Ю.Е., Спокойный В.Г. Об эффективности одного метода рандомизации зеркального спуска в задачах онлайн оптимизации // ЖВМиМФ. Т. 55. № 4. 2015. С. 55–71. [arXiv:1410.3118](https://arxiv.org/abs/1410.3118)
24. Nesterov Y. Primal-dual subgradient methods for convex problems // Math. Program. Ser. B. 2009. V. 120(1). P. 261–283.
25. Поляк Б.Т. Новый метод типа стохастической аппроксимации // Автоматика и телемеханика. 1990. № 7. С. 98–107.
26. Polyak B.T., Juditsky A.B. Acceleration of stochastic approximation by averaging // SIAM J. Control Optim. 1992. V. 30. P. 838–855.
27. 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.
28. Juditsky A., Nemirovski A. First order methods for nonsmooth convex large-scale optimization, I, II. In: Optimization for Machine Learning. Eds. S. Sra, S. Nowozin, S. Wright. MIT Press, 2012.
29. Shapiro A., Dentcheva D., Ruszczynski A. Lecture on stochastic programming. Modeling and theory. MPS-SIAM series on Optimization, 2014.
30. Sridharan K. Learning from an optimization viewpoint. PhD Thesis, Toyota Technological Institute at Chicago, 2011.
31. Границин О.Н., Поляк Б.Т. Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах. М.: Наука, 2003.
32. Spokoiny V. Parametric estimation. Finite sample theory // The Annals of Statistics. 2012. V. 40. № 6. P. 2877–2909. [arXiv:1111.3029v5](https://arxiv.org/abs/1111.3029v5)
33. Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания. М.: Наука, 1977.
34. Rakhlin A., Shamir O., Sridharan K. Making gradient descent optimal for strongly convex stochastic optimization // e-print, 2012. [arXiv:1109.5647](https://arxiv.org/abs/1109.5647)
35. Juditsky A., Nesterov Yu. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // Stoch. System. 2014. V. 4. no. 1. P. 44–80. [arXiv:1401.1792](https://arxiv.org/abs/1401.1792)
36. Hazan E., Kale S. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization // JMLR. 2014. V. 15. P. 2489–2512.
37. Боровков А.А., Боровков К.А. Асимптотический анализ случайных блужданий. Т. 1. Медленно убывающие распределения скачков. М.: Физматлит, 2008.
38. Ким К., Нестеров Ю., Соколов В., Черкасский Б. Эффективные алгоритмы для дифференцирования и задачи экстремали // Экономика и математические методы. 1984. Т. 20. С. 309–318.
39. Евтушенко Ю.Г. Оптимизация и быстрое автоматическое дифференцирование. М.: ВЦ РАН, 2013.
40. Bubeck S., Cesa-Bianchi N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems // Foundation and Trends in Machine Learning. 2012. V. 5. № 1. P. 1–122. <http://www.princeton.edu/~sbubeck/SurveyBCB12.pdf>
41. Agarwal A., Bartlett P.L., Ravikumar P., Wainwright M.J. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization // e-print, 2011. [arXiv:1009.0571](https://arxiv.org/abs/1009.0571)

42. Назин А.В., Поляк Б.Т. Рандомизированный алгоритм нахождения собственного вектора стохастической матрицы с приложением к задаче PageRank // ДАН РАН. 2009. Т. 426. № 6. С. 734–737.
43. Nesterov Y.E. Efficiency of coordinate descent methods on large scale optimization problem // SIAM Journal on Optimization. 2012. V. 22. № 2. P. 341–362.
44. Nesterov Y.E. Subgradient methods for huge-scale optimization problems // Math. Program., Ser. A. 2014 (in print); CORE Discussion Paper 2012/2. 2012.
45. Гасников А.В., Дмитриев Д.Ю. Об эффективных рандомизированных алгоритмах поиска вектора PageRank // ЖВМиМФ. Т. 55. № 3. 2015. С. 355–371. [arXiv:1410.3120](https://arxiv.org/abs/1410.3120)
46. Fercoq O., Richtárik P. Accelerated, parallel and proximal coordinate descent // e-print, 2013. [arXiv:1312.5799](https://arxiv.org/abs/1312.5799)
47. Гасников А.В., Дорн Ю.В., Нестеров Ю.Е., Штирко С.В. О трехстадийной версии модели стационарной динамики транспортных потоков // Математическое моделирование. 2014. Т. 26:6. С. 34–70. [arXiv:1405.7630](https://arxiv.org/abs/1405.7630)
48. Richtárik P. <http://www.maths.ed.ac.uk/~richtarik/>
49. Konecny J., Richtárik P. Semi-Stochastic gradient descent methods // e-print, 2013. [arXiv:1312.1666](https://arxiv.org/abs/1312.1666)
50. Nesterov Yu. Random gradient-free minimization of convex functions // CORE Discussion Paper 2011/1. 2011.
51. Nesterov Y. Smooth minimization of non-smooth function // Math. Program. Ser. A. 2005. V. 103. № 1. P. 127–152.
52. Devolder O., Glineur F., Nesterov Yu. First order methods of smooth convex optimization with inexact oracle // Math. Progr. Ser. A. Accepted. 2013.
53. Devolder O. Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. CORE UCL, PhD thesis, March 2013. http://www.ecore.be/DPs/dp_1327057920.pdf
54. D'Aspermont A. Smooth optimization with approximate gradient // SIAM Journal of Optimization. 2008. V. 19(3). P. 1171–1183.
55. Baes M. Estimate sequence methods: extensions and approximations. IFOR Internal report. ETH Zurich, Switzerland, 2009.
56. Devolder O., Glineur F., Nesterov Yu. First order methods with inexact oracle: the smooth strongly convex case // CORE Discussion Paper 2013/16. 2013.
57. Devolder O., Glineur F., Nesterov Yu. Intermediate gradient methods for smooth convex problems with inexact oracle // CORE Discussion Paper 2013/17. 2013.
58. Ghadimi S., Lan G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework // SIAM J. Optim. 2012. V. 22(4) P. 1469–1492.
59. Ghadimi S., Lan G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: Shrinking procedures and optimal algorithms // SIAM J. Optim. 2013 V. 23(4). P. 2061–2089.
60. Гасников А.В., Девуреченский П.Е. Стохастический промежуточный метод для задач выпуклой оптимизации // ДАН РАН. 2015. (принята к печати) [arXiv:1411.2876](https://arxiv.org/abs/1411.2876)
61. Nesterov Yu., Shikhman V. Convergent subgradient methods for nonsmooth convex minimization. // CORE Discussion Paper 2014/5. 2014.

62. Гасников А.В., Гасникова Е.В., Нестеров Ю.Е., Чепанов Н.С. Об эффективных численных методах решения задач энтропийно-линейного программирования в случае разреженных матриц // e-print, 2014. [arXiv:1410.7719](https://arxiv.org/abs/1410.7719)
63. Lan G. <http://www.ise.ufl.edu/glan/publications/>
64. Spielman D. Algorithms, graph theory, and linear equations in Laplacian matrices // Proc. of the International Congress of Mathematicians. Hyderabad, India, 2010. P. 1–23.
65. Никайдо Х. Выпуклые структуры и математическая экономика. М.: Мир, 1972.
66. Nesterov Yu. Excessive gap technique in nonsmooth convex minimization // SIAM Journal of Optimization. 2005. V. 16. № 1. P. 235–249.
67. Nemirovski A. <http://www2.isye.gatech.edu/~nemirovs/>
68. Juditsky A. <http://ljk.imag.fr/membres/Anatoli.Iouditski/>
69. Nesterov Y., de Palma A. Stationary dynamic solutions in congested transportation Networks: Summary and Perspectives // Networks Spatial Econ. 2003. № 3(3). P. 371–395.
70. Nesterov Yu., Shikhman V. Algorithmic models of market equilibrium // CORE Discussion Paper 2013/66. 2013.
71. Ващенко М.П., Гасников А.В., Молчанов Е.Г., Поспелова Л.Я., Шананин А.А. Вычислимые модели и численные методы для анализа тарифной политики железнодорожных грузоперевозок. М.: ВЦ РАН, 2014. [arXiv:1501.02205](https://arxiv.org/abs/1501.02205)
72. Гасников А.В. Заметка об эффективной вычислимости конкурентных равновесий в транспортно-экономических моделях // Математическое моделирование. 2015. (принята к печати) [arXiv:1410.3123](https://arxiv.org/abs/1410.3123)
73. Nemirovski A., Onn S., Rothblum U.G. Accuracy certificates for computational problems with convex structure // Mathematics of Operation Research. 2010. V. 35. № 1. P. 52–78.
74. Nesterov Yu. New primal-dual subgradient methods for convex optimization problems with functional constraints // International Workshop “Optimization and Statistical Learning”. January 11–16. France, Les Houches, 2015. <http://lear.inrialpes.fr/workshop/osl2015/program.html>
75. Магарил-Ильяев Г.Г., Тихомиров В.М. Выпуклый анализ и его приложения. М.: УРСС, 2011.
76. Zheng Q., Richtárik P., Zhang T. Randomized dual coordinate ascent with arbitrary sampling // e-print, 2014. [arXiv:1411.5873](https://arxiv.org/abs/1411.5873)
77. Guzman C., Nemirovski A. On lower complexity bounds for large-scale smooth convex optimization // Journal of Complexity. 2015. [arXiv:1307.5001](https://arxiv.org/abs/1307.5001)
78. Harchaoui Z., Juditsky A., Nemirovski A. Conditional gradient algorithms for norm-regularized smooth convex optimization // Math. Program. Ser. B. 2015. http://www2.isye.gatech.edu/~nemirovs/ccg_revised_apr02.pdf
79. Nesterov Yu. Universal gradient methods for convex optimization problems // CORE Discussion Paper 2013/63. 2013.
80. Гасников А.В., Двуреченский П.Е., Камзолов Д.И. Градиентные и прямые методы с неточным оракулом для задач стохастической оптимизации // Труды СДСР-2014. Екатеринбург, 2015. (принята к печати) [arXiv:1502.06259](https://arxiv.org/abs/1502.06259)
81. Nesterov Yu. Gradient methods for minimizing composite functions // Math. Prog. 2013. V. 140. № 1. P. 125–161.
82. Горнов А.Ю. Вычислительные технологии решения задач оптимального управления. Новосибирск: Наука, 2009.

83. *Lugosi G., Cesa-Bianchi N.* Prediction, learning and games. New York: Cambridge University Press, 2006.
84. *Shalev-Shwartz S.* Online learning and online convex optimization // Foundation and Trends in Machine Learning. 2011. V. 4. № 2. P. 107–194.
<http://www.cs.huji.ac.il/~shais/papers/OLsurvey.pdf>
85. *Rakhlin A., Sridharan K.* Statistical Learning Theory and Sequential Prediction // e-print, 2015. http://stat.wharton.upenn.edu/~rakhlin/book_draft.pdf
86. *Hazan E.* Introduction to online convex optimization // e-print, 2015.
<http://ocobook.cs.princeton.edu/OCObook.pdf>
87. Гасников А.В., Лагуновская А.А., Усманова И.Н., Федоренко Ф.А. Безградиентные прокс-методы с неточным оракулом для негладких задач выпуклой стохастической оптимизации на симплексе // Автоматика и телемеханика. 2015. [arXiv:1412.3890](https://arxiv.org/abs/1412.3890)
88. Гасников А.В., Крымова Е.А., Усманова И.Н., Федоренко Ф.А. Стохастическая онлайн оптимизация. Одноточечные и многоточечные бандиты. Сильно выпуклый случай // Проб. перед. информ. 2016.
89. *Duchi J.C., Jordan M.I., Wainwright M.J., Wibisono A.* Optimal rates for zero-order convex optimization: the power of two function evaluations // e-print, 2014. [arXiv:1312.2139](https://arxiv.org/abs/1312.2139)
90. Поляк Б.Т., Цыбаков А.Б. Оптимальные порядки точности поисковых алгоритмов стохастической оптимизации // Проб. перед. информ. 1990. Т. 26. № 2. С. 45–53.
91. *Spall J.C.* Introduction to stochastic search and optimization: estimation, simulation and control. Wiley, 2003.
92. *Agarwal A., Dekel O., Xiao L.* Optimal algorithms for online convex optimization with multi-point bandit feedback // Proceedings of 23-d Annual Conference on Learning Theory. 2010. P. 28–40.
93. *Ledoux M.* Concentration of measure phenomenon. Providence, RI, Amer. Math. Soc., 2001 (Math. Surveys Monogr. V. 89).
94. *Flaxman A.D., Kalai A.T., McMahan H.B.* Online convex optimization in the bandit setting: gradient descent without a gradient // e-print, 2004. [arXiv:cs/0408007](https://arxiv.org/abs/cs/0408007)
95. *Dempe S.* Foundations of bilevel programming. Dordrecht: Kluwer Academic Publishers, 2002.
96. *Dvurechensky P., Gasnikov A., Zhukovskii M.* Random gradient-free methods for random walk based web page ranking functions learning // e-print, 2014. [arXiv:1411.4282](https://arxiv.org/abs/1411.4282)