

Репорт 1. Предварительный анализ

Некоторые начальные рассуждения о поведении методов на малой задаче

Dmitry Pasechnyuk

О характере сходимости методов

- Методы можно разделить на две категории: локальные и глобальные. Локальные методы: Nadam, SVRG, Adam, Adamax, Yogi, Apollo, A2GradExp, DiffGrad; глобальные методы: Adadelata, AdaBound, Lamb, NovoGrad. Локальные методы сходятся к глубокой точке в устойчивом (за счёт стохастической сходимости) локальном минимуме, в то время как глобальные методы особенно эффективно используют овражный характер моментной схемы и достигают минимума, более близкого к глобальному / находящегося в области притяжения глобального минимума. Это можно видеть из графиков изменения функции потерь, вычисленной на валидационной выборке: для локальных методов график существенно немонокотонен, что говорит о низкой обобщающей способности, в то время как для глобальных методов наблюдается устойчивое убывание (за исключением тех из них, что сразу достигают достаточно малых значений функции потерь и далее не могут значительно улучшить их, ориентируясь только лишь на обучающую выборку).
- При проверке на тестовой выборке разброс показателей качества решений, полученных методами, одинаково велик в обеих категориях и не позволяет выделить одну из них как более эффективную. Тем не менее, видны значительные изменения в ранжировании методов, в направлении улучшения показателей глобальных методов.
- Среди локальных методов наилучшими оказались Adamax, Yogi и DiffGrad: достигаемые ими показатели качества превосходят их значения для всех прочих рассматриваемых методов.
- Достаточно эффективны оказались глобальные методы AdaBound и Lamb. Помимо превосходящей обобщающей способности и сравнительно неплохих показателей качества, метод AdaBound демонстрирует наибольший темп сходимости, причём сходимость практически не затухает при увеличении числа итераций. Это также указывает на то, что прочие рассматриваемые методы сходятся к локальному минимуму прежде, чем достигают области регулярных минимумов. Метод же Lamb использует

свои глобализующие свойства на начальном этапе оптимизации, быстро спускаясь к широкому оптимуму, и попав в него естественно замедляет сходимость.

- Наилучшим глобальным методом является RAdam, максимально использующий свои глобализующие свойства, благодаря чему подойти к глобальному минимуму ему удаётся за время, на порядок меньшее соответствующей характерной временной протяжённости для других методов. В то же время, вместе с углублением в найденный локальный минимум (точнее, судя по всему, в локальную ложбину в квази-оптимальной долине – артефакт используемой эмпирической аппроксимации функции риска) обобщающая способность метода значительно ухудшается. Вероятно, это ждёт все методы на этапе, когда они подойдут к оптимуму настолько близко, так что это более вопрос Statistical learning'a, нежели оптимизации.
- Смотря на графики, можно выделить три пучка методов, различающиеся характером сходимости соответствующих методов. Первый пучок (SVRG, A2GradExp, Apollo, возможно Adadelta) затухает на начальном этапе оптимизации, сходясь к локальному экстремуму – он соответствует локальным методам. Второй пучок (Nadam, Adam, Yogi, Adamax и проч.) быстро подходит к области притяжения глубокого (квази-глобального) минимума, и затухает по мере приближения к области регулярности – соответствующие методы обладают некоторым глобализующим потенциалом, за счёт которого преодолевают серию неглубоких локальных минимумов на начальном участке траектории (эффект овражности или, возможно, уравнивания острых минимумов "минных полей"). Наиболее же интересен в отношении дальнейшей разработки третий пучок методов (AdaBound, Lamb, NovoGrad), не относящихся визуально ни к первому, ни ко второму пучку, и не удовлетворяющих общей тенденции к затуханию: при медленной сходимости вначале, сходимость на более поздних этапах может сохраняться или даже ускоряться, – такие методы обладают наиболее гибкими глобализующими способностями, и их свойства могут быть использованы для построения комбинированных методов, тем или иным образом ведомых exploring-процедурой из числа этих методов (в том числе с использованием ансамблирования по типу Ветрова). Мнемоническое правило для вычисления "особых" методов: локальное постоянство производной, локальная линейность (в общем гометрическом смысле) сходимости, нехарактерная для выпуклой оптимизации, периодическое ускорение темпа сходимости.

Об областях применимости методов

- Одними из наименее удовлетворительных показателей качества выделяется метод SVRG. Это можно объяснить тем, что метод изначально более предназначен для выпуклых задач, в том числе его теоретический анализ наивно использует условие

Ферма, а на практике практически отсутствуют инструменты глобализации его как обучающей процедуры. Гипотеза состоит в том, что метод SVRG в данном случае достаточно быстро попал в область притяжения локального экстремума, причём характеристическое свойство метода – редукция дисперсии – сыграло в данном случае роль препятствия к возможному освобождению из паразитного минимума за счёт стохастической динамики. Решение состоит в том, чтобы либо использовать метод SVR как *finalizer* в мультиметодной схеме, либо отказаться от его использования в принципе в том случае если глобальный экстремум не обладает свойством локальной выпуклости (для нейронных сетей типично невыполнение этого условия).

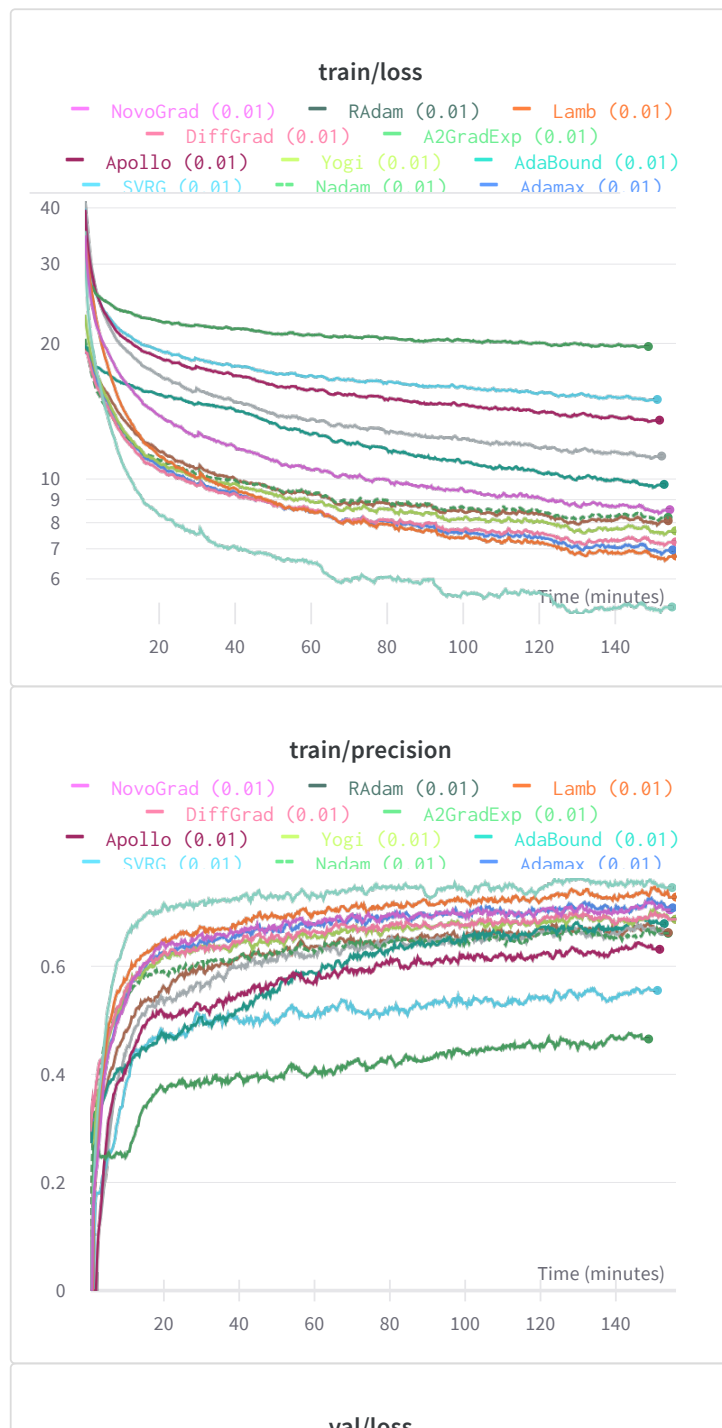
- Метод A2GradExp, будучи ускоренным методом, также как и SVRG, несколько привязан к выпуклой структуре функционала и потому имеет тенденцию к застреванию в ближайшем локальном минимуме. Наиболее вероятно, что и дальнейшее его использование не будет оправдано.
- Метод Apollo реализует принцип квази-ньютоновской аппроксимации в стохастической постановке, и как и многие квази-ньютоновские методы имеет потенциал в использовании долгострочной истории сходимости, эксплуатируя тем самым инерционность рассматриваемой модели. Низкая эффективность метода в данных тестах может быть объяснена тем, что для попадания в область притяжения квази-глобального минимума методу необходимо преодолеть ряд локальных минимумов, как бы "пролетая" над ними – овражность в данном случае имеет больше седловой характер, и модель в этой области не обладает инерционностью. Возможно, в области глобального минимума квази-ньютоновская схема будет показывать себя более лучшим образом.
- Одним из наименее успешных методов при оценке на обучающем наборе (то есть непосредственно по величине минимизируемого функционала) стал метод Adadelta: при сохранении достаточного темпа сходимости, из-за его производительности на начальной эпохе его итоговый результат оказался неудовлетворительным. Естественным решением проблемы является использование метода вместе с каким-нибудь из прочих методов, используемым в качестве *starter*, так чтобы начальная точка уже принадлежала некоторому котловану.

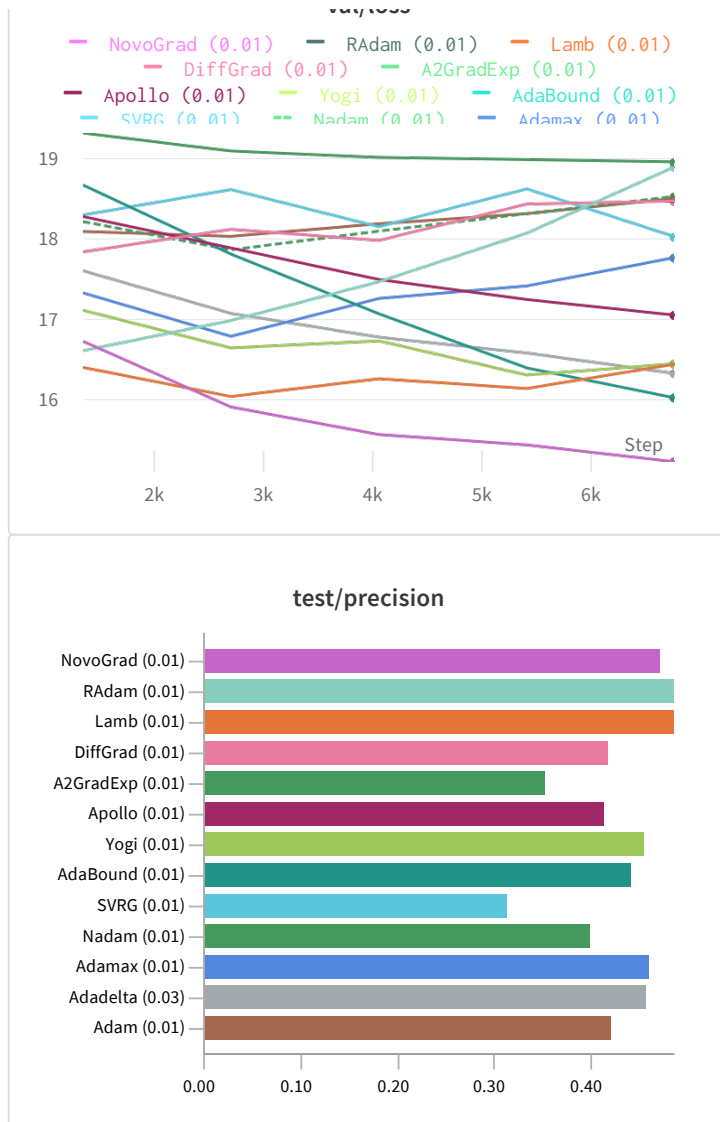
Выводы

- Наиболее перспективными среди рассмотренных являются методы AdaBound, Lamb, NovoGrad и RAdam, ввиду специфической способности сохранения ими темпа сходимости / не-следования принципу затухания / наличия сильного глобализующего потенциала.
- Методы редукции дисперсии неэффективны на начальном этапе решения задачи, такж

как ускоренные и квази-ньютоновские методы.


- Остаётся невыясненным вопрос о возможности использования полноградиентных методов с малым батчем (ожидаются БОЛЬШИЕ вычислительные мощности, вариант LBFGS)
- Остаётся невыясненным вопрос о полезности использования информации второго порядка (вариант Adahessian).
- Остаётся невыясненным вопрос об ускорении сходимости, достигаемом с помощью предобуславливания задачи (вариант Shampoo).





Run set



Created with  on Weights & Biases.

<https://wandb.ai/dmivilensky/code2seq-java-small/reports/-1---VmIldzo1MDk2MDY>