Hyperfast Second-Order Method

Nikita Dudorov
Optimization Class Project. MIPT

Introduction

Let us observe the problem of unconstrained convex optimization for functions with Lipschitz-continuous third derivative. Applying methods, which use derivatives up to order 3, one has the best possible convergence rate in function value $O(k^{-5})$ [1]. A significant difficulty for third-order methods is computation of the third derivative. In [2] Nesterov has shown that in tensor methods one can approximate third-order information by gradients and preserve the high convergence rate $O(k^{-4})$. Near-optimal third-order algorithm with convergence rate $\tilde{O}(k^{-5})$ is proposed in [3]. Both results are combined in [4] to get an algorithm of order 2 and convergence rate which, up to a logarithmic factor, coincide the optimal rate of third-order methods. The purpose of the project is to implement this so called Hyperfast Second-Order Method and compare it with the fast gradient method.

Auxiliary Problem

Consider the augmented Taylor polynomial of convex function $f : R^n \to R$ whose p-th derivative is Lipschitz:

$$\Omega_{x,p,H_p}(y) = f(x) + \sum_{i=1}^{p} \frac{1}{i!} D^i f(x)[y-x]^i + \frac{H_p}{p!}\|y-x\|^{p+1} \to \min_{y \in R^n}$$

It could be shown that this problem is convex if $H_p \geq L_p$.

Accelerated Taylor Descent

Proposed in [3] algorithm:

ATD

Initialize $A_0 = 0$, $x_0 = y_0 = 0$
for $k = 0, 1, \ldots$ **do**
  Compute $\lambda_{k+1} > 0$ and y_{k+1} such that
  $$\frac{1}{2} \leq \lambda_{k+1} \frac{L_p \|y_{k+1} - \tilde{x}_k\|}{(p-1)!} \leq \frac{p}{p+1}$$
  where
  $$y_{k+1} = \operatorname*{argmin}_y \Omega_{\tilde{x}_k, p, L_p}(y)$$
  and
  $$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1} A_k}}{2}, \quad A_{k+1} = A_k + a_{k+1},$$
  $$\tilde{x}_k = \frac{A_k}{A_{k+1}} y_k + \frac{a_{k+1}}{A_{k+1}} x_k$$
  Update $x_{k+1} = x_k - a_{k+1} \nabla f(y_k)$
end for
return y_k

ensures

$$f(y_k) - f(x_*) \leq \frac{C(p) L_p \|x_*\|^{p+1}}{k^{(3p+1)/2}}$$

Particularly, the convergence rate is $\tilde{O}(k^{-5})$ for $p = 3$.

Inexact auxiliary problem solution

Let us call for $\gamma \in [0; 1)$ any point from the set

$$N^{\gamma}_{p,H_p}(x) = \{T : \|\nabla \Omega_{x,p,H_p}(T)\|_* \leq \gamma \|\nabla f(T)\|_*\}$$

an inexact solution of the auxiliary problem. Note that if $\gamma = 0$, $N^0_{p,H_p}(x)$ is the exact solution. According to [4], taken $p = 3, \gamma = \frac{1}{6}, H_3 = \frac{3}{2} L_3$ one could satisfy the requirements of **ATD** and at each iteration find a point from the set $N^{1/6}_{3,3L_3/2}(\tilde{x_k})$ instead of solving the auxiliary problem.

Approximate Gradients

As shown in [2], one could compute approximate value of $\nabla \Omega(y)$ using

$$g^{\tau}_x(y) = \frac{1}{\tau^2}(\nabla f(x + \tau(y-x)) + \nabla f(x - \tau(y-x)) - 2\nabla f(x))$$

and get accuracy

$$\|g^{\tau}_x(y) - D^3 f(x)[y-x]^2\|_* \leq \frac{\tau}{3} L_3 \|y-x\|^3$$

Consequently, if the total error of approximate $\nabla \Omega(y)$ is δ, $y \in N^{1/6}_{3,3L_3/2}(x)$ if

$$\|g(y)\|_* \leq \frac{1}{6}\|\nabla f(y)\|_* - \delta$$

where

$$g(y) = \nabla f(x) + D^2 f(x)[y-x] + \frac{1}{2} g^{\tau}_x(y) + L_3 \|y-x\|^2 (y-x)$$

Bregman-Distance Gradient Method

According to [2], [4] one could find a point from $N^{1/6}_{3,3L_3/2}(x)$ with the following algorithm:

BDGM

Set $\tau = \frac{3\delta}{8(2+\sqrt{2})\|\nabla f(x)\|_*}$, $y_0 = x$
Set
$$\rho(y) = \frac{1}{2} D^2 f(x)[y-x]^2 + L_3 \frac{\|y-x\|^4}{4}$$
Set
$$\beta_\rho(x, y) = \rho(y) - \rho(x) - \langle \nabla \rho(x), y - x \rangle$$
for $k = 0, 1, \ldots$ **do**
  Compute $g(y_k)$
  if
  $$\|g(y_k)\|_* \leq \frac{1}{6}\|\nabla f(x)\|_* - \delta$$
  then
    Stop
  else
  $$y_{k+1} = \operatorname*{argmin}_y \left(\langle g(y_k), y - y_k \rangle + 2\left(1 + \frac{1}{\sqrt{2}}\right) \beta_\rho(y_k, y) \right)$$
  end if
end for
return y_k

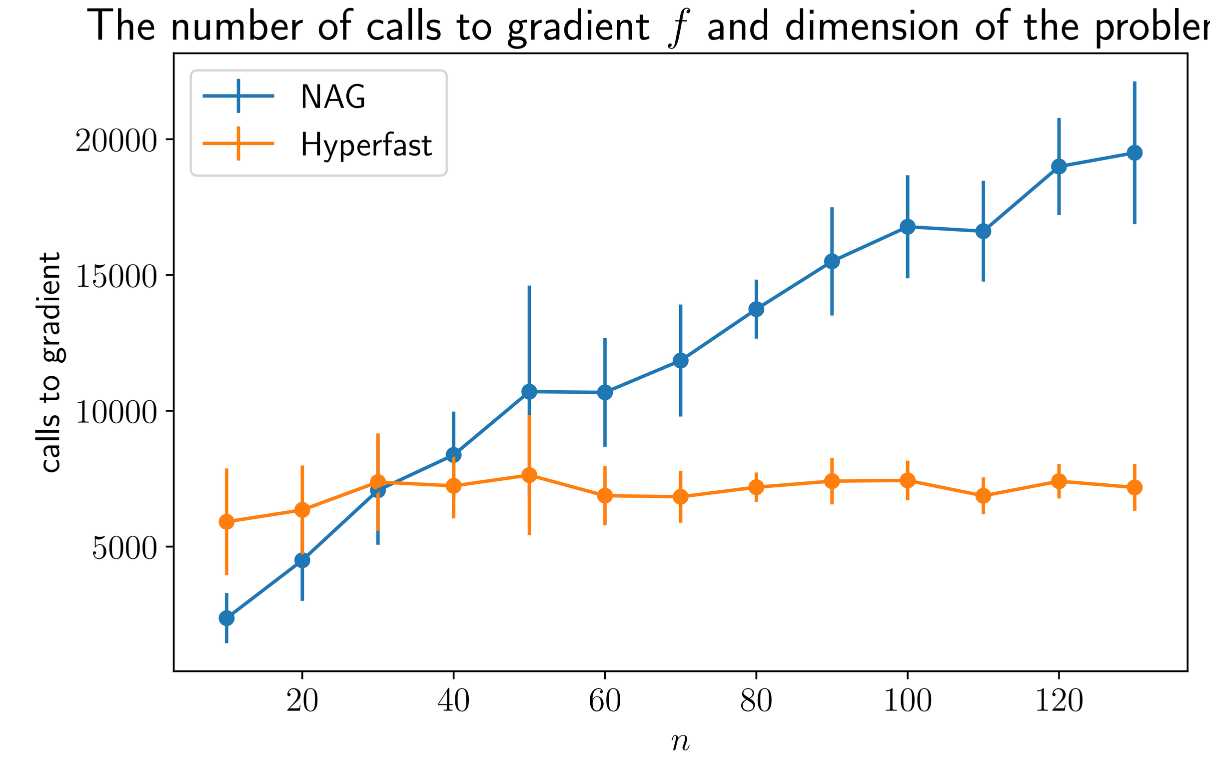
Hyperfast Second-Order Method

Now it is clear that simple combination of **ATD** and **BDGM** which is to change
ATD, STEP 3: $L_3 \to \frac{3}{2} L_3$
ATD, STEP 4: $y_{k+1} \in N^{1/6}_{3,3L_3/2}(\tilde{x_k})$ computed by **BDGM**
gives **Hyperfast** method, which has order 2 and convergence rate $\tilde{O}(k^{-5})$

Experiments and Results

Solve: $f(x) = \sum_{i=1}^{n} \log(1 + \exp a_i^T x) \to \min_{x \in R^n}$
with accuracy: $|f(x) - f(x_*)| \leq \varepsilon = 10^{-4}$
by **Hyperfast** and Nesterov Accelerated Gradient **NAG** [5] and compute the number of calls to gradient f for both methods. Code.

The number of calls to gradient f and dimension of the problem n



Conclusion

In sense of oracle complexity, the Hyperfast Second-Order Method turned out to be more efficient than accelerated gradient descent, whose convergence rate is $O(k^{-2})$. This result is consistent with the theory. However, **Hyperfast** requires more time as it does a lot more intermediate computations. In particular, about 90% of time is spent on solving the auxiliary problem in **BDGM** by **NAG**. It offers hope for **Hyperfast** to replace fast gradient methods in some cases, if one finds more effective way of solving that problem.

References

[1] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 2019.

[2] Yurii Nesterov. Superfast second-order methods for unconstrained convex optimization. to appear, 2020.

[3] Sbastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near-optimal method for highly smooth convex optimization, 2018.

[4] Dmitry Kamzolov and Alexander Gasnikov. Near-optimal hyperfast second-order method for convex optimization and its sliding. to appear, 2020.

[5] Stephen Boyd Weijie Su and Emmanuel J. Cand'es. A differential equation for modeling nesterovs accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 2016.