Adaptive Negative Curvature Descent with Applications in Non-convex Optimization

Mingrui Liu†, Zhe Li†, Xiaoyu Wang‡, Jinfeng Yi♮, Tianbao Yang†

†Department of Computer Science, The University of Iowa ‡Intellifusion ♮JD AI Research

Background and Main Contributions

Background: Negative curvature descent (NCD) needs to approximate the smallest eigen-value of the Hessian matrix with a sufficient precision in order to achieve a sufficiently accurate second-order stationary solution, which is computationally expensive.

Contributions:

1. Proposed a variant of NCD, i.e., adaptive Negative Curvauture Descent to allow an adaptive error dependent on the current gradient's magnitude in approximating the smallest eigen-value of the Hessian, which is able to reduce the overall complexity in computing negative curvature without sacrificing the iteration complexity
2. Verified the practical effectiveness of the proposed algorithms by three experiments (cubic regularization, regularized non-linear least square, and one hidden layer neural network)

Problem of Interest

Problem: Consider

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \tag{1}$$

Assumption:

1. $f(\mathbf{x})$ has L_1-Lipschitz continuous gradient and L_2-Lipschitz continuous Hessian.
2. Given an initial solution $\mathbf{x}_0$, there exists $\Delta < \infty$ such that $f(\mathbf{x}_0) - f(\mathbf{x}_*) \leq \Delta$, where $\mathbf{x}_*$ denotes the global minimum of (1);
3. if $f(\mathbf{x})$ is a stochastic objective, we assume each random function $f(\mathbf{x};\xi)$ is twice differentiable and has L_1-Lipschitz continuous gradient and L_2-Lipschitz continuous Hessian, and its stochastic gradient has exponential tail behavior, i.e., $\mathbb{E}[\exp(\|\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{x})\|^2/G^2)] \leq \exp(1)$ holds for any $\mathbf{x}\in\mathbb{R}^d$;
4. a Hessian-vector product can be computed in $O(d)$ time.

Goal: find an approximate second-order stationary point with:

$$\|\nabla f(\mathbf{x})\| \leq \epsilon_1, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\epsilon_2, \tag{2}$$

In our paper, we assume $\epsilon_2 = \epsilon_1^{\alpha}$. Define $f_{\mathcal{S}}(\mathbf{x}) = \frac{1}{|\mathcal{S}|}\sum_{\xi\in\mathcal{S}} f(\mathbf{x};\xi)$, $\mathbf{g}(\mathbf{x}) = \frac{1}{|\mathcal{S}|}\sum_{\xi\in\mathcal{S}} \nabla f(\mathbf{x};\xi)$, $H_{\mathcal{S}}(\mathbf{x}) = \frac{1}{|\mathcal{S}|}\sum_{\xi\in\mathcal{S}} \nabla^2 f(\mathbf{x};\xi)$.

Negative Curvature Search: Assume there exists an algorithm that can compute a unit-length negative curvature direction $\mathbf{v}\in\mathbb{R}^d$ of a function $f(\mathbf{x})$ satisfying

$$\lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq \mathbf{v}^\top\nabla^2 f(\mathbf{x})\mathbf{v} - \varepsilon \tag{3}$$

with high probability $1-\delta$. We refer to such an algorithm as $\text{NCS}(f, \mathbf{x}, \varepsilon, \delta)$ and denote its time complexity by $T_n(f, \varepsilon, \delta, d)$.

Adaptive Negative Curvature Step

Algorithm 1 AdaNCD$^{\text{det}}(\mathbf{x}, \alpha, \delta, \nabla f(\mathbf{x}))$

Apply $\text{NCS}(f, \mathbf{x}, \frac{\max(\epsilon_2, \|\nabla f(\mathbf{x})\|^{\alpha})}{2}, \delta)$ to find a unit vector $\mathbf{v}$ satisfying (3)
if $\frac{2(-\mathbf{v}^\top\nabla^2 f(\mathbf{x})\mathbf{v})^3}{3L_2^2} > \frac{\|\nabla f(\mathbf{x})\|^2}{2L_1}$ **then**
  $\mathbf{x}^+ = \mathbf{x} - \frac{2|\mathbf{v}^\top\nabla^2 f(\mathbf{x})\mathbf{v}|}{L_2}\text{sign}(\mathbf{v}^\top\nabla f(\mathbf{x}))\mathbf{v}$,
else
  $\mathbf{x}^+ = \mathbf{x} - \frac{1}{L_1}\nabla f(\mathbf{x})$
end if
return $\mathbf{x}^+, \mathbf{v}$

Algorithm 2 AdaNCD$^{\text{mb}}(\mathbf{x}, \alpha, \delta, \mathcal{S}, \mathbf{g}(\mathbf{x}))$

Apply $\text{NCS}(f_{\mathcal{S}}, \mathbf{x}, \frac{\max(\epsilon_2, \|\mathbf{g}(\mathbf{x})\|^{\alpha})}{2}, \delta)$ to find a unit vector $\mathbf{v}$ satisfying (3)
if $\frac{2(-\mathbf{v}^\top H_{\mathcal{S}}(\mathbf{x})\mathbf{v})^3}{3L_2^2} - \frac{\epsilon_2|\mathbf{v}^\top H_{\mathcal{S}}(\mathbf{x})\mathbf{v}|^2}{6L_2^2} > \frac{\|\mathbf{g}(\mathbf{x})\|^2}{4L_1} - \frac{\epsilon'^2}{L_1}$ **then**
  $\mathbf{x}^+ = \mathbf{x} - \frac{2|\mathbf{v}^\top H_{\mathcal{S}}(\mathbf{x})\mathbf{v}|}{L_2} z\mathbf{v}$, $\Pr(z = \pm 1) = 1/2$
else
  $\mathbf{x}^+ = \mathbf{x} - \frac{1}{L_1}\mathbf{g}(\mathbf{x})$
end if
return $\mathbf{x}^+, \mathbf{v}$

Theoretical Gurantee of AdaNCD Step

Deterministic Objective When $\mathbf{v}^\top\nabla^2 f(\mathbf{x})\mathbf{v} \leq 0$, the Algorithm 1 (AdaNCD$^{\text{det}}$) provides a guarantee that

$$f(\mathbf{x}) - f(\mathbf{x}^+) \geq \max\left(\frac{2|\mathbf{v}^\top\nabla^2 f(\mathbf{x})\mathbf{v}|^3}{3L_2^2}, \frac{\|\nabla f(\mathbf{x})\|^2}{2L_1}\right)$$

Stochastic Objective Assume $\|H_{\mathcal{S}}(\mathbf{x}) - \nabla^2 f(\mathbf{x})\|_2 \leq \epsilon_2/12$ and $\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \epsilon'$ hold (with high probability). When $\mathbf{v}^\top H_{\mathcal{S}}(\mathbf{x})\mathbf{v} \leq 0$, the Algorithm 2 (AdaNCD$^{\text{mb}}$) provides a guarantee (with high probability) that

$$f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x}^+)] \geq \max\left\{\frac{2(-\mathbf{v}^\top H_{\mathcal{S}}(\mathbf{x})\mathbf{v})^3}{3L_2^2} - \frac{\epsilon_2|\mathbf{v}^\top H_{\mathcal{S}}(\mathbf{x})\mathbf{v}|^2}{6L_2^2}, \frac{\|\mathbf{g}(\mathbf{x})\|^2}{4L_1} - \frac{\epsilon'^2}{L_1}\right\}$$

If $\mathbf{v}^\top H_{\mathcal{S}}(\mathbf{x})\mathbf{v} \leq -\epsilon_2/2$, we have

$$f(\mathbf{x}) - f(\mathbf{x}^+) \geq \max\left(\frac{\epsilon_2^3}{24L_2^2}, \frac{\|\mathbf{g}(\mathbf{x})\|^2}{4L_1} - \frac{\epsilon'^2}{L_1}\right)$$

Simple Adaptive Algorithms with Second-order Convergence

Algorithm 3 AdaNCG: $(\mathbf{x}_0, \epsilon_1, \alpha, \delta)$

$\mathbf{x}_1 = \mathbf{x}_0$, $\epsilon_2 = \epsilon_1^{\alpha}$
$\delta' = \delta/(1 + \max\left(\frac{12L_2^2}{\epsilon_2^3}, \frac{2L_1}{\epsilon_1^2}\right)\Delta)$,
for $j = 1, 2, \ldots,$ **do**
  $(\mathbf{x}_{j+1}, \mathbf{v}_j) =$ AdaNCD$^{\text{det}}(\mathbf{x}_j, \alpha, \delta', \nabla f(\mathbf{x}))$
  if $\mathbf{v}_j^\top\nabla^2 f(\mathbf{x}_j)\mathbf{v}_j > -\frac{\epsilon_2}{2}$ and $\|\nabla f(\mathbf{x}_j)\| \leq \epsilon_1$ **then**
    return $\mathbf{x}_j$
  end if
end for

Algorithm 4 S-AdaNCG: $(\mathbf{x}_0, \epsilon_1, \alpha, \delta)$

$\mathbf{x}_1 = \mathbf{x}_0$, $\epsilon_2 = \epsilon_1^{\alpha}$, $\delta' = \delta/\widetilde{O}(\epsilon_1^{-2}, \epsilon_2^{-3})$
for $j = 1, 2, \ldots,$ **do**
  Generate two random sets $\mathcal{S}_1, \mathcal{S}_2$
  let $\mathbf{g}(\mathbf{x}_j) = \frac{1}{|\mathcal{S}_1|}\sum_{\xi\in\mathcal{S}_1}\nabla f(\mathbf{x};\xi)$
  $(\mathbf{x}_{j+1}, \mathbf{v}_j) =$ AdaNCD$^{\text{mb}}(\mathbf{x}_j, \alpha, \delta', \mathcal{S}_2, \mathbf{g}(\mathbf{x}_j))$
  if $\mathbf{v}_j^\top H_{\mathcal{S}_2}(\mathbf{x}_j)\mathbf{v}_j > -\epsilon_2/2$ and $\|\mathbf{g}(\mathbf{x}_j)\| \leq \epsilon_1$ **then**
    return $\mathbf{x}_j$
  end if
end for

Deterministic Objective For any $\alpha \in (0, 1]$, the AdaNCG algorithm terminates at iteration j_* for some

$$j_* \leq 1 + \max\left(\frac{12L_2^2}{\epsilon_1^{3\alpha}}, \frac{2L_1}{\epsilon_1^2}\right)(f(\mathbf{x}_1) - f(\mathbf{x}_{j_*})) \leq 1 + \max\left(\frac{12L_2^2}{\epsilon_1^{3\alpha}}, \frac{2L_1}{\epsilon_1^2}\right)\Delta, \tag{4}$$

with $\|\nabla f(\mathbf{x}_{j_*})\| \leq \epsilon_1$, and with probability at least $1-\delta$, $\lambda_{\min}(\nabla^2 f(\mathbf{x}_{j_*})) \geq -\epsilon_1^{\alpha}$. Furthermore, the j-th iteration requires time a complexity of $T_n(f, \max(\epsilon_1^{\alpha}, \|\nabla f(\mathbf{x}_j)\|^{\alpha}), \delta', d)$.

Stochastic Objective Set $|\mathcal{S}_1| = \frac{32G^2}{\epsilon_1^2}(1 + 3\log(\frac{2}{\delta'}))$ and $|\mathcal{S}_2| = \frac{9216L_1^2}{\epsilon_2^2}\log(\frac{4d}{\delta'})$. With probability $1-\delta$, the S-AdaNCG algorithm terminates at some iteration $j_* = \widetilde{O}(\max\left(\frac{1}{\epsilon_2^3}, \frac{1}{\epsilon_1^2}\right))$ and upon termination it holds that $\|\nabla f(\mathbf{x}_{j_*})\| \leq 2\epsilon_1$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x}_{j_*})) \geq -2\epsilon_2$ with probability $1-3\delta$. Furthermore, the worst-case time complexity of S-AdaNCG is given by $\widetilde{O}\left(\max\left(\frac{1}{\epsilon_2^3}, \frac{1}{\epsilon_1^2}\right)\left(\frac{d}{\epsilon_1^2} + T_n(f_{\mathcal{S}_2}, \epsilon_2, \delta', d)\right)\right)$.

Adaptive Algorithms with State-of-the-Art Complexities

Algorithm 5 AdaNCG$^+$: $(\mathbf{x}_0, \epsilon_1, \alpha, \delta)$

$\delta' = \delta/\left(\left\lceil 1 + \Delta\left(\frac{\max(12L_2^2, 2L_1)}{\epsilon_2^3} + \frac{2\sqrt{10}L_2}{\epsilon_1\epsilon_2}\right)\right\rceil\right)$
for $k = 1, 2, \ldots,$ **do**
  $\widehat{\mathbf{x}}_k =$ AdaNCG$(\mathbf{x}_k, \epsilon_1^{3\alpha/2}, \frac{2}{3}, \delta')$
  if $\|\nabla f(\widehat{\mathbf{x}}_k)\| \leq \epsilon_1$ **then**
    return $\widehat{\mathbf{x}}_k$
  else
    $f_k(\mathbf{x}) = f(\mathbf{x}) + L_1([\|\mathbf{x} - \widehat{\mathbf{x}}_k\| - \epsilon_2/L_2]_+)^2$
    $\mathbf{x}_{k+1} =$ Almost-Cvx-AGD$(f_j, \widehat{\mathbf{x}}_k, \frac{\epsilon_1}{2}, 3\epsilon_2, 5L_1)$
  end if
end for

Algorithm 6 AdaNCD-SCSG: $(\mathbf{x}_0, \epsilon_1, \alpha, b, \delta)$

Input: $\mathbf{x}_0$, ϵ_1, α, δ
for $j = 1, 2, \ldots,$ **do**
  Generate three random sets $\mathcal{S}, \mathcal{S}_1, \mathcal{S}_2$
  $\mathbf{y}_j =$ SCSG-Epoch$(\mathbf{x}_j, \mathcal{S}, b)$
  let $\mathbf{g}(\mathbf{y}_j) = \nabla f_{\mathcal{S}_1}(\mathbf{x};\xi)$
  $(\mathbf{x}_{j+1}, \mathbf{v}_j) =$ AdaNCD$^{\text{mb}}(\mathbf{y}_j, \alpha, \delta, \mathcal{S}_2, \mathbf{g}(\mathbf{y}_j))$
  if $\mathbf{v}_j^\top H_{\mathcal{S}_2}(\mathbf{y}_j)\mathbf{v}_j > -\epsilon_2/2$ and $\|\mathbf{g}(\mathbf{y}_j)\| \leq \epsilon_1$ **then**
    return $\mathbf{y}_j$
  end if
end for

Deterministic Objective With probability at least $1-\delta$, the Algorithm AdaNCG$^+$ returns a vector $\widehat{\mathbf{x}}_k$ such that $\|\nabla f(\widehat{\mathbf{x}}_k)\| \leq \epsilon_1$ and $\lambda_{\min}(\nabla^2 f(\widehat{\mathbf{x}}_k)) \geq -\epsilon_2$ with at most $O\left(\frac{1}{\epsilon_2^3} + \frac{1}{\epsilon_1\epsilon_2}\right)$ AdaNCD steps in AdaNCG and $\widetilde{O}\left[\left(\frac{1}{\epsilon_2^{7/2}} + \frac{1}{\epsilon_1\epsilon_2^{3/2}}\right) + \frac{\epsilon_2^{1/2}}{\epsilon_1^2}\right]$ gradient steps in Almost-Convex-AGD, and each step j within AdaNCG$^+$ requires time of $T_n(f, \max(\epsilon_2, \|\nabla f(\mathbf{x}_j)\|^{2/3})^{1/2}, \delta', d)$, and the worse-case time complexity of AdaNCG$^+$ is $\widetilde{O}\left(\left(\frac{d}{\epsilon_1\epsilon_2^{3/2}} + \frac{d}{\epsilon_2^{7/2}}\right) + \frac{d\epsilon_2^{1/2}}{\epsilon_1^2}\right)$ when using the Lanczos method for NCS.

Stochastic Objective Suppose $|\mathcal{S}| = \widetilde{O}(\max(1/\epsilon_1^2, 1/(\epsilon_2^{9/2}b^{1/2})))$, $|\mathcal{S}_1| = \widetilde{O}(1/\epsilon_1^2)$ and $|\mathcal{S}_2| = \widetilde{O}(1/\epsilon_2^2)$. With high probability, the Algorithm AdaNCD-SCSG returns a vector $\mathbf{y}_j$ such that $\|\nabla f(\mathbf{y}_j)\| \leq 2\epsilon_1$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x}_j)) \geq -2\epsilon_2$ with at most $\widetilde{O}\left(\frac{b^{1/3}}{\epsilon_1^{4/3}} + \frac{1}{\epsilon_2^3}\right)$ calls of SCSG-Epoch and AdaNCD$^{\text{mb}}$. When $\epsilon_1 = \epsilon$, $\epsilon_2 = \sqrt{\epsilon}$, then by choosing $b = \frac{1}{\sqrt{\epsilon}}$, the complexity is $\widetilde{O}(\frac{d}{\epsilon^{3.5}})$.

Experimental Results

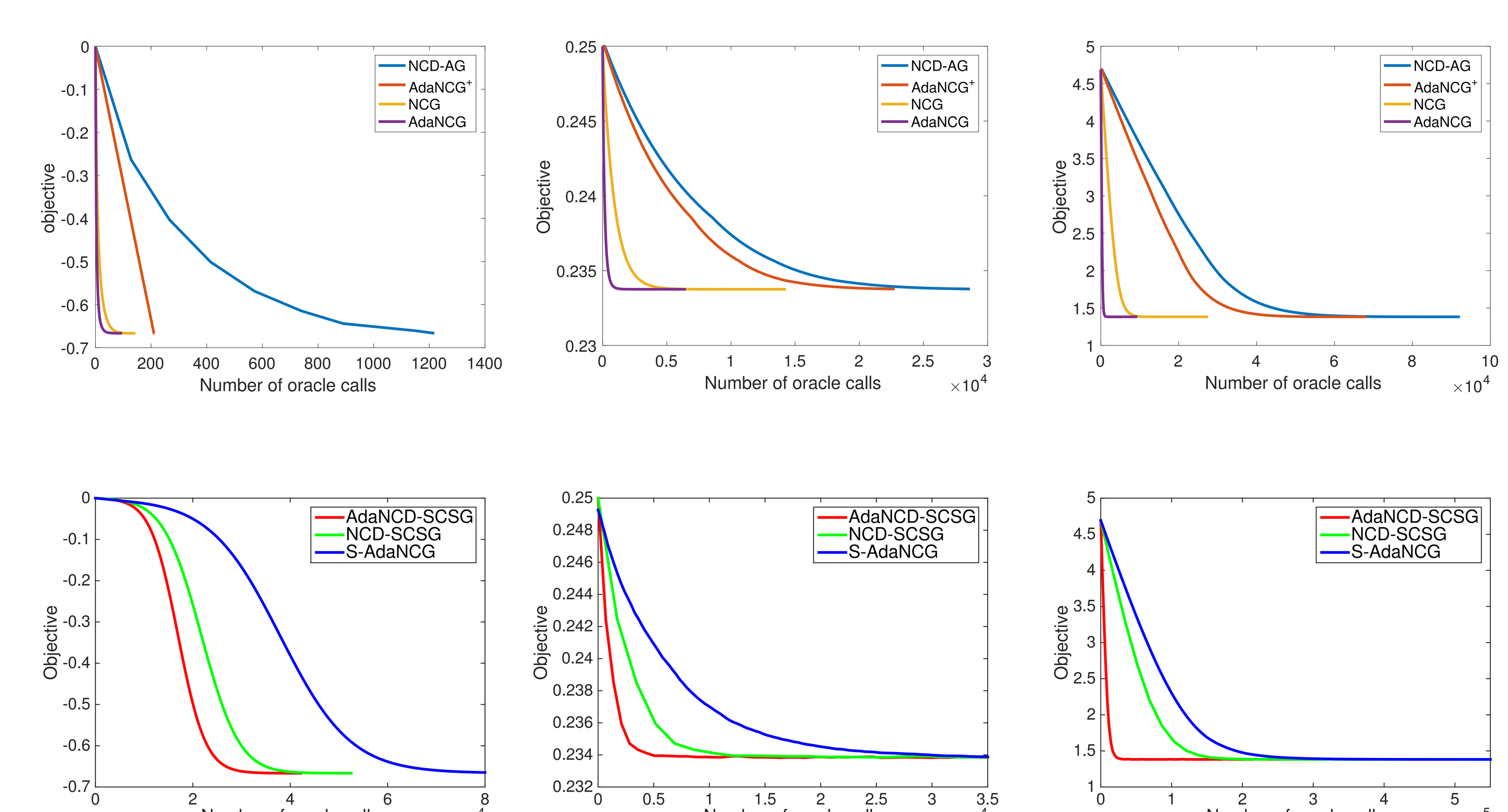

Figure: Comparison of deterministic algorithms (upper) and stochastic algorithms (lower) for solving cubic regularization, regularized nonlinear least square, and NN (from left to right).