A randomized algorithm for nonconvex minimization with inexact evaluations and complexity guarantees

Shuyao Li [1]
Joint work with Stephen Wright [1]

SIAM Conference on Optimization
June 1, 2023

[1] University of Wisconsin—Madison

Problem Setup

Find an approximate second-order stationary point (SOSP) x^* of

$$\min_{x \in \mathbb{R}^d} \quad f(x).$$

- (ϵ_g, ϵ_H)-approximate SOSP (we assume no ϵ_g and ϵ_H coupling): $\|\nabla f(x^*)\| \leq \epsilon_g, \quad \lambda_{\min}\left(\nabla^2 f(x^*)\right) \geq -\epsilon_H.$
- f has L-Lipschitz gradient and M-Lipschitz Hessian
- f is bounded below by $\bar{f} > -\infty$.
- Inexact evaluations at iterate x_k:
  - Inexact gradient g_k such that $\|g_k - \nabla f(x_k)\| \leq \frac{1}{3}\max\{\epsilon_g, \|g_k\|\}$
  - Inexact Hessian $\mathbf{H}_k$ such that $\|\mathbf{H}_k - \nabla^2 f(x_k)\|_{\text{op}} \leq \frac{2}{9}\epsilon_H$
  - Only need Hessian for a fraction of iterations
  - No function evaluation $f(x_k)$ is needed
- More general than mini-batching in stochastic optimization

Basic algorithm with exact evaluations

Algorithm 1: Wright and Recht 2022[Section 3.6]

if $\|\nabla f(x_k)\| > \epsilon_g$ **then**
    `// gradient step`
    $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$
else if $\lambda_k := \lambda_{\min}(\nabla^2 f(x_k)) < -\epsilon_H$ **then**
    `// negative curvature step`
    $p_k \leftarrow$ unit minimum eigenvector of $\nabla^2 f(x_k)$ with $\nabla f(x_k)^\top p_k \leq 0$
    $x_{k+1} = x_k + \frac{2\epsilon_H}{M} p_k$
else
    return x_k

Basic algorithm with exact evaluations: complexity

- Gradient descent analysis is standard

$$f(x_{k+1}) \leq f(x_k) - \frac{\epsilon_g^2}{2L}$$

- Negative curvature step:

$$\begin{aligned} f(x_{k+1}) &= f(x_k + \frac{2\epsilon_H}{M} p_k) \\ &\leq f(x_k) + 2\frac{\epsilon_H}{M} \underbrace{\nabla f(x_k)^\top p_k}_{\leq 0} + \frac{1}{2} \cdot \frac{4\epsilon_H^2}{M^2} \underbrace{p_k^\top \nabla^2 f(x_k) p_k}_{< -\epsilon_H} + \frac{M}{6} \cdot \frac{8\epsilon_H^3}{M^3} \\ &\leq f(x_k) - \frac{2\epsilon_H^3}{3M^2} \end{aligned}$$

- Complexity guarantee: Algorithm 1 terminates at an (ϵ_g, ϵ_H)-approximate SOSP in at most

$$\frac{f(x_0) - \bar{f}}{\min\left(\frac{\epsilon_g^2}{2L}, \frac{2\epsilon_H^3}{3M^2}\right)} \text{ iterations.}$$

Algorithm with inexact evaluations

Inexact gradient g_k such that $\|g_k - \nabla f(x_k)\| \leq \frac{1}{3}\max\{\epsilon_g, \|g_k\|\}$

Inexact Hessian $\mathbf{H}_k$ such that $\|\mathbf{H}_k - \nabla^2 f(x_k)\| \leq \frac{2}{9}\epsilon_H$

Algorithm 2: Our algorithm

if $\|g_k\| > \epsilon_g$ **then**

```
// gradient step
```

$x_{k+1} = x_k - \frac{1}{L} g_k$

else if $\hat{\lambda}_k := \lambda_{\min}(\mathbf{H}_k) < -\epsilon_H$ **then**

```
// negative curvature step
```

$\hat{p}_k \leftarrow$ unit minimum eigenvector of $\mathbf{H}_k$

Draw $\sigma_k \leftarrow \pm 1$ with probability $\frac{1}{2}$

$x_{k+1} = x_k + \frac{2\epsilon_H}{M} \sigma_k \hat{p}_k$

else

return x_k

Complexity Guarantee

Theorem

- *If Algorithm 2 terminates and returns x_n, then x_n is an $(\frac{4}{3}\epsilon_g, \frac{4}{3}\epsilon_H)$-approximate SOSP.*
- **Expected:** *Let N denote the iteration at which Algorithm 2 terminates. Then $N < \infty$ with probability one and*

$$\mathbb{E}N \leq \frac{f(x_0) - \bar{f}}{C_\epsilon}, \qquad C_\epsilon := \min\left(\frac{\epsilon_g^2}{6L}, \frac{2\epsilon_H^3}{9M^2}\right)$$

Same complexity as the deterministic algorithm with exact evaluations.

- **High-Probability:** *Algorithm 2 terminates after n iterations with probability $1 - \delta$, for*

$$n = O\left(\frac{f(x_0) - \bar{f}}{C_\epsilon} + \frac{1}{\tau^2}\left(\frac{ML\epsilon_g}{\epsilon_H^3}\right)^{1+\tau} \log\left(\frac{1}{\delta}\right)\right),$$

where we can choose τ to be a small constant at the expense of a large constant factor

Interpreting the high-probability complexity guarantee

$$n = \tilde{O}\Bigg(\overbrace{\frac{f(x_0) - \bar{f}}{C_\epsilon}}^{\text{Expected}} + \overbrace{\frac{1}{\tau^2}\Big(\frac{ML\epsilon_g}{\epsilon_H^3}\Big)^{1+\tau}}^{\text{High probability correction}}\Bigg)$$

$$C_\epsilon = \min\left(\frac{\epsilon_g^2}{6L}, \frac{2\epsilon_H^3}{9M^2}\right)$$

Corollary

$\epsilon_H = \sqrt{\epsilon_g M}$ *Choosing* $\tau = 1$ *gives* $n = \tilde{O}(\frac{1}{\epsilon_g^2})$.

ϵ_g *and* ϵ_H *satisfy* $\frac{\epsilon_g^2}{6L} = \frac{2\epsilon_H^3}{9M^2}$ *Choosing* $\tau = 1$ *gives* $n = \tilde{O}(\frac{1}{\epsilon_g^2})$.

No coupling between ϵ_g and ϵ_H required

Previous work (Yao et al. 2022) considered the same general inexact settings, but

- Can only handle $\epsilon_H = O(\sqrt{\epsilon_g})$— "strong coupling"
- Analyze the cubic to choose a stepsize

$$
\begin{aligned}
f(x_{k+1}) &= f(x_k + \frac{2\alpha_k}{M}\hat{p}_k) \\
&\leq f(x_k) + 2\frac{\alpha_k}{M}\nabla f(x_k)^\top \hat{p}_k + \frac{1}{2}\cdot\frac{4\alpha_k^2}{M^2}\hat{p}_k^\top \nabla^2 f(x_k)\hat{p}_k + \frac{M}{6}\cdot\frac{8\alpha_k^3}{M^3}
\end{aligned}
$$

- Lead to worse (stricter) gradient inexactness tolerance

No ϵ_g, ϵ_H coupling required: matrix factorization

An example where breaking the strong coupling between ϵ_g and ϵ_H leads to relaxed requirements on gradient accuracy while attaining the same solution quality.

$$f(\mathbf{U}) = \frac{1}{2} \left\| \mathbf{U}\mathbf{U}^\top - \mathbf{M}^* \right\|_F^2$$

- $\mathbf{M}^* \in \mathbb{R}^{d \times d}$ is the unknown symmetric and positive semidefinite.
- $\mathrm{rank}(\mathbf{M}^*) = r < d$. The variable is $\mathbf{U} \in \mathbb{R}^{d \times r}$.
- $\sigma_1^\star$—the largest singular value of $\mathbf{M}^*$
  $\sigma_r^\star$—the smallest nonzero singular value of $\mathbf{M}^*$.

No ϵ_g, ϵ_H coupling required: matrix factorization

Properties of f (Jin et al. 2017):

- All local minima are global minima — $\mathcal{X}^\star$
- $(\frac{1}{24}{\sigma_r^\star}^{3/2}, \frac{1}{3}\sigma_r^\star)$-approximate SOSP is $\frac{1}{3}{\sigma_r^\star}^{1/2}$-close to $\mathcal{X}^\star$
- f satisfies local regularity condition in $\frac{1}{3}{\sigma_r^\star}^{1/2}$-neighborhood of $\mathcal{X}^\star$
  - gradient descent converges linearly inside this neighborhood to an arbitrarily accurate solution.
  - Algorithm 2 gets us to this neighborhood.
- For any $\Gamma > \sigma_1^\star$, inside the region $\{\mathbf{U} : \|\mathbf{U}\|_{\text{op}}^2 < \Gamma\}$, $f(\cdot)$ is
  - $L = 16\Gamma$-gradient Lipschitz
  - $M = 24\Gamma^{\frac{1}{2}}$-Hessian Lipschitz.

No ϵ_g, ϵ_H coupling required: matrix factorization

Need $(\frac{1}{24}\sigma_r^{\star 3/2}, \frac{1}{3}\sigma_r^\star)$-approximate SOSP. Hessian Lipschitzness $M = 24\Gamma^{1/2}$.

Let $\kappa = \Gamma/\sigma_r^\star$

- Our work: $\epsilon_g \sim \sigma_r^{\star 3/2} \quad \epsilon_H \sim \sigma_r^\star$
- Previous work (Yao et al. 2022): $\epsilon_H = \sqrt{\epsilon_g M}$
  $\epsilon_g \lesssim \sigma_r^{\star 3/2}, \sqrt{\epsilon_g M} \lesssim \sigma_r^\star \implies \epsilon_g \sim \frac{\sigma_r^{\star 3/2}}{\sqrt{\kappa}} \quad \epsilon_H \sim \sigma_r^\star.$

$\|g_k - \nabla f(x_k)\| \lesssim \epsilon_g$: Decoupling allows us to tolerate more error in the approximate gradient.

Concrete scenario in which only inexact evaluations are available: robust low-rank matrix sensing with Gaussian design (upcoming work)

- Sensing matrices $\mathbf{A}_i \in R^{d\times d}$ have i.i.d. standard Gaussian entries
- Measurements $y_i = \langle \mathbf{A}_i, \mathbf{M}^* \rangle = \mathrm{tr}(\mathbf{A}_i^\top \mathbf{M}^*)$
- $f_i(\mathbf{U}) = (\langle \mathbf{U}\mathbf{U}^\top, \mathbf{A}_i \rangle - y_i)^2 \implies \mathbb{E} f_i(\mathbf{U}) = f(\mathbf{U})$
- A fraction of $\{(\mathbf{A}_i, y_i)\}$ are arbitrarily corrupted

Relative gradient inexactness

Inexact gradient g_k True gradient $\nabla f(x_k)$

- Previous works: $\|g_k - \nabla f(x_k)\| \leq \frac{1}{3}\epsilon_g$
- Our work: $\|g_k - \nabla f(x_k)\| \leq \frac{1}{3}\max\{\epsilon_g, \|g_k\|\}$
  Alternatively $\|g_k - \nabla f(x_k)\| \leq \frac{1}{4}\max\{\epsilon_g, \|\nabla f(x_k)\|\}$

Our algorithm is the first that tolerates **relative** gradient inexactness for **second-order guarantee** to the best of our knowledge

(Tolerating relative gradient inexactness for first-order guarantee is well-studied in, *e.g.*, Paquette and Scheinberg 2020)

Relative gradient inexactness: finite-sum subsampling

Theorem

For a given $x \in \mathbb{R}^d$*, suppose there is an upper bound* $G(x)$ *such that* $\|\nabla f_i(x)\|_2 \leq G(x) < \infty$ *for all sample indices* i*.*

For any given $\xi \in (0,1)$*, if* $|S_g(x)| \geq \Omega\left(\frac{G(x)}{\max\{\epsilon_g, \|\nabla f(x)\|\}} \log(\xi)\right)^2$ *where* $S_g(x)$ *is with-replacement sub-sampling indices, then for* $g(x) := \frac{1}{|S_g(x)|} \sum_{i \in S_g(x)} \nabla f_i(x)$*, we have*

$$\mathbb{P}\left(\|\nabla f(x) - g(x)\|_2 \leq \frac{1}{3} \max\{\epsilon_g, \|g(x)\|\}\right) \geq 1 - \xi. \qquad (1)$$

Cartis and Scheinberg 2018 has a similar relative gradient estimate, and they proposed an adaptive scheme for choosing $|S_g(x)|$ based on it.

Analysis (Expectation result)

- Gradient step:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{6L}\epsilon_g^2$$

- Negative curvature step:

$$f(x_{k+1}) \leq f(x_k) - \frac{2\epsilon_H^3}{9M^2} + 2\frac{\alpha_k}{M}\nabla f(x_k)^\top \sigma_k \hat{p}_k$$

Combining:

$$\mathbb{E}\left[f(x_{k+1})|x_k\right] \leq f(x_k) - \min\left(\frac{\epsilon_g^2}{6L}, \frac{2\epsilon_H^3}{9M^2}\right) = f(x_k) - C_\epsilon$$

Hence $M_k := f(x_k) + kC_\epsilon$ is a supermartingale, *i.e.*, $\mathbb{E}(M_{k+1}\,|\mathcal{G}_k) \leq M_k$

Our algorithm stops at iteration $N \implies N$ is a stopping time.

Optional stopping theorem: $\mathbb{E}M_N \leq \mathbb{E}M_0$

Analysis (Expectation result)

$M_k := f(x_k) + kC_\epsilon \quad \mathbb{E}M_N \leq \mathbb{E}M_0$

$\mathbb{E}M_N = \mathbb{E}f(x_N) + \mathbb{E}N \cdot C_\epsilon \geq \bar{f} + \mathbb{E}N \cdot C_\epsilon$
$\mathbb{E}M_0 = f(x_0)$

Hence

$$\mathbb{E}N \leq \frac{f(x_0) - \bar{f}}{C_\epsilon} = \frac{f(x_0) - \bar{f}}{\min\left(\frac{\epsilon_g^2}{6L}, \frac{2\epsilon_H^3}{9M^2}\right)}$$

Analysis (High probability result)

Analysis is much more complicated.

Markov inequality: with probability at least δ, it holds that $N \leq \frac{f(x_0)-\bar{f}}{\delta C_\epsilon}$.

Our complexity bound has only **logarithmic** dependence on δ.

Main elements of the analysis:

- Bound the function value increase of "wrong" negative curvature steps
- Cannot have too many "wrong" steps, by Azuma-Hoeffding's inequality
- Use the descent lemma from gradient descent to offset wrong negative curvature steps

Summary

- Finding SOSPs using inexact gradients and Hessians
- Simple short step method, no function value evaluation needed
- "Flip a coin" to determine the sign of negative curvature steps
- Complexity obtained for expected and high probability runtime: comparable to deterministic algorithm with exact evaluations
- Requires no coupling between ϵ_g and ϵ_H (helpful for some problems, *e.g.*, robust low-rank matrix sensing)
- Relative gradient inexactness condition
- Motivated by applications to robust low-rank matrix sensing

References I

Cartis, Coralia and Katya Scheinberg (2018). “Global convergence rate analysis of unconstrained optimization methods based on probabilistic models”. In: *Mathematical Programming* 169, pp. 337–375.

Jin, Chi et al. (2017). “How to Escape Saddle Points Efficiently”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, pp. 1724–1732.

Paquette, Courtney and Katya Scheinberg (2020). “A stochastic line search method with expected complexity analysis”. In: *SIAM J. Optim.* 30.1, pp. 349–376. ISSN: 1052-6234. DOI: 10.1137/18M1216250. URL: https://doi.org/10.1137/18M1216250.

References II

Wright, Stephen J. and Benjamin Recht (2022). *Optimization for data analysis*. Cambridge University Press, Cambridge, pp. x+227. ISBN: 978-1-316-51898-4. DOI: 10.1017/9781009004282. URL: https://doi.org/10.1017/9781009004282.

Yao, Zhewei et al. (Aug. 2022). "Inexact Newton-CG algorithms with complexity guarantees". In: *IMA Journal of Numerical Analysis*. ISSN: 0272-4979. DOI: 10.1093/imanum/drac043. URL: https://doi.org/10.1093/imanum/drac043.