First-order Stochastic Algorithms for Escaping From Saddle Points in Almost Linear Time

Yi Xu†, Rong Jin‡, Tianbao Yang†

†Computer Science Department, University of Iowa, Iowa City, IA, USA
‡Machine Intelligence Technology, Alibaba Group, Bellevue, WA, USA

Stochastic Non-convex Optimization Problem

The optimization problem of interest:

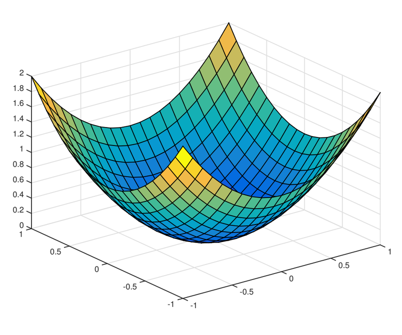
$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \mathrm{E}_\xi[f(\mathbf{x};\xi)], \quad (1)$$

where ξ is a random variable, $F(\mathbf{x})$ and $f(\mathbf{x};\xi)$ are non-convex. Let denote by $\mathbf{x}_*$ the global minimum of (1).
We make the following assumptions:

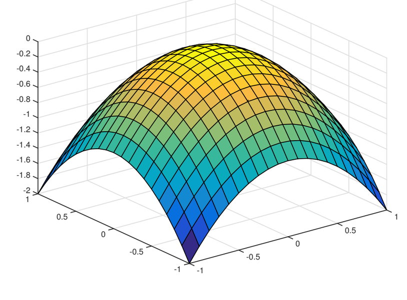
- every $f(\mathbf{x};\xi)$ is twice differentiable, and it has L_1-Lipschitz continuous gradient and L_2-Lipschitz continuous Hessian.
- given an initial point $\mathbf{x}_0$, $\exists\Delta < \infty$ s.t. $F(\mathbf{x}_0) - F(\mathbf{x}_*) \le \Delta$.
- $\exists G > 0$ s.t. $\mathbb{E}[\exp(\|\nabla f(\mathbf{x};\xi) - \nabla F(\mathbf{x})\|^2/G^2)] \le \exp(1)$.

Introduction

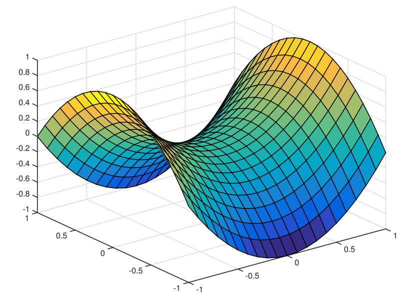
- Non-convex optimization is challenging: in general, finding global minimum of non-convex optimization is NP-hard.
- Finding critical points is relatively easy: first-order stationary point (FSP) $\|\nabla F(\mathbf{x})\| = 0$.
  - First-order necessary condition of local minimum
  - Iteration complexity of SGD [5,8]: $O(1/\epsilon^4)$ for finding ϵ-FSP, $\mathrm{E}[\|\nabla F(\mathbf{x})\|_2^2] \le \epsilon^2$.
  - Improved iteration complexity of SCSG (variance reduction based) [7]: $O\left(1/\epsilon^{10/3}\right)$.



local min: $\nabla^2 F(\mathbf{x}) \succeq 0$



local max: $\nabla^2 F(\mathbf{x}) \prec 0$



saddle point: $\lambda_{\min}(\nabla^2 F(\mathbf{x})) < 0$

- To find **second-order stationary points (SSP)**:

$$\|\nabla F(\mathbf{x})\|_2 = 0, \lambda_{\min}(\nabla^2 F(\mathbf{x})) \ge 0.$$

  - Second-order necessary condition of local minimizer.
  - For strict saddle functions: FSP is either a local minimizer or a non-degenerate saddle point $\Longrightarrow$ SSP is local minimum.
- Goal: finding an **approximate local minimum** by using first-order methods

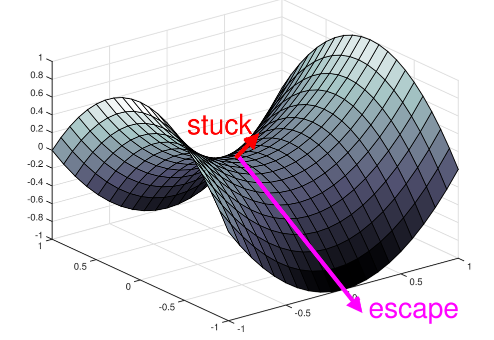
$$(\epsilon,\gamma)-\mathrm{SSP}: \quad \|\nabla F(\mathbf{x})\|_2 \le \epsilon, \quad \lambda_{\min}(\nabla^2 F(\mathbf{x})) \ge -\gamma$$

Related Work

- Adding Isotropic Noise: Noisy SGD [4], SGLD [9]
  - Time complexity: $\tilde{O}(d^p/\epsilon^4)$, where $p \ge 5$
  - not practical for high-dimensional optimization problems
- Using full gradient (FG) and isotropic noise: Perturbed GD [6]
  - Add perturbation around a saddle point $\widetilde{\mathbf{x}}_t = \mathbf{x}_t + n_t$, take GD from $\widetilde{\mathbf{x}}_t$
  - Time complexity: almost linear dependence on d
- Using Hessian-vector product (HVP): Natasha2 [2]
  - can take $O(d)$ runtime for particular problems with special structures
- Using both FG and HVP [1, 3]

Escape from Saddle Points

- Motivation: How to Escape from Saddle Points?



- $F(\mathbf{x}+\Delta) \approx F(\mathbf{x}) + \Delta^\top\nabla F(\mathbf{x}) + \frac{L}{2}\Delta^\top\nabla^2 F(\mathbf{x})\Delta$
- Saddle points have zero gradient, i.e., $\nabla F(\mathbf{x}) = 0$
- Non-degenerate Hessian, i.e. $\lambda_{\min}(\nabla^2 F(\mathbf{x})) < 0$
- Negative eigenvector is a direction of escaping

- Definition: Suppose $\lambda_{\min}(\nabla^2 F(\mathbf{x})) \le -\gamma$, a direction $\mathbf{v}\in\mathbb{R}^d$ is called negative curvature (NC) direction if it satisfies ($c > 0$ is a constant)

$$\mathbf{v}^\top\nabla^2 F(\mathbf{x})\mathbf{v} \le -c\gamma \text{ and } \|\mathbf{v}\| = 1$$

- Finding NC: second-order methods, e.g., Power method and Lanczos method

$$\mathbf{v}_0 = \mathbf{n}, \quad \text{// isotropic noise}$$
$$\mathbf{v}_{t+1} = (I - \eta\nabla^2 F(\mathbf{x}))\mathbf{v}_t \quad \text{//Power method}$$

NEON: NEgative curvature Originated from Noise

- **NEON is a new perspective of noise perturbation**
  - Inspired by Perturbed GD [6]: around a saddle point $\mathbf{x}$

$$\mathbf{x}_0 = \mathbf{x} + \mathbf{e}, \text{ nosie } \mathbf{e} \text{ is from sphere of a Euclidean ball}$$
$$\mathbf{x}_\tau = \mathbf{x}_{\tau-1} - \eta\nabla F(\mathbf{x}_{\tau-1}), \tau = 1,\ldots,$$

  - An Equivalent Sequence: let $\mathbf{u}_\tau = \mathbf{x}_\tau - \mathbf{x}$

$$\mathbf{u}_\tau = \mathbf{u}_{\tau-1} - \eta\nabla F(\mathbf{u}_{\tau-1}+\mathbf{x}) \approx \mathbf{u}_{\tau-1} - \eta(\nabla F(\mathbf{u}_{\tau-1}+\mathbf{x}) - \nabla F(\mathbf{x}))$$
$$\approx \mathbf{u}_{\tau-1} - \eta\nabla^2 F(\mathbf{x})\mathbf{u}_{\tau-1} = (I - \eta\nabla^2 F(\mathbf{x}))\mathbf{u}_{\tau-1}$$

  - Around saddle point: PGD $\approx$ Power method
  - NEON update: starting with a random noise $\mathbf{u}_0$, the recurrence:

$$\mathbf{u}_\tau = \mathbf{u}_{\tau-1} - \eta(\nabla F(\mathbf{x}+\mathbf{u}_{\tau-1}) - \nabla F(\mathbf{x})), \tau = 1,\ldots$$

Algorithm 1 NEON($f, \mathbf{x}, t, \mathcal{F}, r$)

```
Input: f, x, t, F, r
Generate u_0 randomly from S_r^d
for τ = 0, ..., t do
  u_{τ+1} = u_τ − η(∇f(x + u_τ) − ∇f(x))
end for
if min_{i∈[t+1],‖u_i‖≤U} f(x + u_i) − f(x) − ∇f(x)^⊤ u_i ≤ −2.5F then
  return u_{τ'}, τ' = arg min_{i∈[t+1],‖u_i‖≤U} f̂_x(u_i)
else
  return 0
end if
```

Main Result 1 (NEON)

Theorem 1. Suppose $\mathbf{x}$ satisfies $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \le -\gamma$. With $\mathcal{F} = \widetilde{O}(\gamma^3)$ $r = \widetilde{O}(\gamma^2)$, $U = \widetilde{O}(\gamma)$, then after $t = \widetilde{O}\left(\frac{1}{\gamma}\right)$ iterations, with high probability $1-\delta$ NEON returns $\mathbf{u} \ne 0$ such that

$$\mathbf{v}^\top\nabla^2 f(\mathbf{x})\mathbf{v} \le -\widetilde{\Omega}(\gamma), \quad \mathbf{v} = \mathbf{u}/\|\mathbf{u}\|.$$

- $\mathbf{v}$ is a NC of $\nabla^2 f(\mathbf{x})$; if NEON returns 0, then $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -\gamma$ with high probability.
- Iteration complexity of NEON is Similar to the Power method
- NEON can find a NC at any point $\mathbf{x}$ whose Hessian has a negative eigen-value regardless close to a saddle point or not

NEON⁺: Accelerated NEON

- NEON is essentially an application of GD to decrease $\hat{f}_{\mathbf{x}}(\mathbf{u})$:

$$\hat{f}_{\mathbf{x}}(\mathbf{u}) = f(\mathbf{x}+\mathbf{u}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top\mathbf{u}.$$

  - Lipschitz continuous Hessian: $\frac{1}{2}\mathbf{u}^\top\nabla^2 f(\mathbf{x})\mathbf{u} \le \hat{f}(\mathbf{u}) + \frac{L_2}{6}\|\mathbf{u}\|^3$.
- Use Nesterov's Accelerated Gradient to decrease $\hat{f}_{\mathbf{x}}(\mathbf{u})$:

$$\mathbf{y}_{\tau+1} = \mathbf{u}_\tau - \eta\nabla\hat{f}_{\mathbf{x}}(\mathbf{u}_\tau), \quad \mathbf{u}_{\tau+1} = \mathbf{y}_{\tau+1} + \zeta(\mathbf{y}_{\tau+1} - \mathbf{y}_\tau)$$

Main Result 2 (NEON⁺)

Theorem 2. Suppose $\mathbf{x}$ satisfies $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \le -\gamma$. With $\mathcal{F} = \widetilde{O}(\gamma^3)$ $r = \widetilde{O}(\gamma^2)$, $U = \widetilde{O}(\gamma)$, momentum parameter $\zeta = 1 - \sqrt{\eta\gamma}$, then after $t = \widetilde{O}\left(\frac{1}{\sqrt{\gamma}}\right)$ iterations, with high probability $1-\delta$ NEON⁺ returns $\mathbf{u} \ne 0$ such that

$$\mathbf{v}^\top\nabla^2 f(\mathbf{x})\mathbf{v} \le -\widetilde{\Omega}(\gamma), \quad \mathbf{v} = \mathbf{u}/\|\mathbf{u}\|.$$

- $\mathbf{v}$ is a NC of $\nabla^2 f(\mathbf{x})$; if NEON returns 0, then $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -\gamma$ with high probability.
- Matches the iteration complexity of Lanczos Method

Stochastic NEON

- Challenge: not easy evaluate gradient of $F(\mathbf{x}) = \mathrm{E}_{\mathbf{x}}[f(\mathbf{x};\xi)]$ exactly
- Resort to mini-batching technique:

$$F_{\mathcal{S}}(\mathbf{x}) = \frac{1}{|\mathcal{S}|}\sum_{\xi\in\mathcal{S}} f(\mathbf{x};\xi), \text{ where } \mathcal{S} = \{\xi_1,\ldots,\xi_m\}$$

- Find an approximate NC $\mathbf{u}_{\mathcal{S}}$ by applying NEON/NEON⁺ to $F_{\mathcal{S}}(\mathbf{x})$

Main Result 3 (Stochastic NEON)

Theorem 3. Let mini-batch size $m = \widetilde{O}(1/\gamma^2)$, then with high probability

$$\mathbf{v}_{\mathcal{S}}^\top\nabla^2 F(\mathbf{x})\mathbf{v}_{\mathcal{S}} \le -\widetilde{\Omega}(\gamma), \quad \mathbf{v}_{\mathcal{S}} = \mathbf{u}_{\mathcal{S}}/\|\mathbf{u}_{\mathcal{S}}\|.$$

NEON and NEON⁺ terminate with a total complexity of $\widetilde{O}(1/\gamma^3)$ and $\widetilde{O}(1/\gamma^{2.5})$, respectively.

First-order Stochastic Algorithms based on NEON

- NEON-$\mathcal{A}$: a framework for promoting $\mathcal{A}$ for finding a SSP based on the proposed stochastic NEON
- Assume $\mathcal{A}$ is a stochastic algorithm that is guaranteed to find a FSP, e.g.,
  - SGD, Stochastic Heavy-ball Method, Stochastic Nesterov's Accelerated Gradient Method
  - SCSG, SVRG

Algorithm 2 NEON-$\mathcal{A}$

```
for j = 1, 2, ..., do
  Running updates of A(x_j)
  if first-order condition not met then
    Take A's output as x_{j+1}
  else
    Update x_{j+1} with a NC direction found by Stochastic NEON
  end if
end for
```

Table: Comparisons of First-order Stochastic Algorithms for achieving an $(\epsilon,\sqrt{\epsilon})$-SSP, where T_h denotes the runtime of stochastic HVP and T_g denotes the runtime of SG.

Algorithm	Target	Time Complexity
Noisy SGD [4]	$(\epsilon,\epsilon^{1/2})$-SSP	$\widetilde{O}(T_g d^p\epsilon^{-4})$, $p \ge 4$
SGLD [9]	$(\epsilon,\epsilon^{1/2})$-SSP	$\widetilde{O}(T_g d^p\epsilon^{-4})$, $p \ge 4$
Natasha2 [2]	$(\epsilon,\epsilon^{1/2})$-SSP	$\widetilde{O}(T_g\epsilon^{-3.5} + T_h\epsilon^{-2.5})$
NEON-SGD, NEON-SM **(this work)**	$(\epsilon,\epsilon^{1/2})$-SSP	$\widetilde{O}(T_g\epsilon^{-4})$
NEON-SCSG **(this work)**	$(\epsilon,\epsilon^{1/2})$-SSP	$\widetilde{O}(T_g\epsilon^{-3.5})$
NEON-Natasha **(this work)**	$(\epsilon,\epsilon^{1/2})$-SSP	$\widetilde{O}(T_g\epsilon^{-3.5})$
NEON-SVRG **(this work)** (finite sum)	$(\epsilon,\epsilon^{1/2})$-SSP	$\widetilde{O}\left(T_g\left(n^{2/3}\epsilon^{-2} + n\epsilon^{-1.5} + \epsilon^{-2.75}\right)\right)$

Conclusions

- Proposed novel first-order procedures to extract NC from a Hessian matrix
- Develop a general framework of first-order stochastic algorithms with a second-order convergence guarantee
- First result of first-order stochastic algorithm with almost linear time complexity for finding SSP

References

1. N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima faster than gradient descent. In STOC, pages 1195-1199, 2017.
2. Z. Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. CoRR, /abs/1708.08694, 2017.
3. Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for non-convex optimization. SIAM Journal on Optimization, 28(2):1751-1772, 2018.
4. R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In COLT, pages 797-842, 2015.
5. S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4):2341-2368, 2013.
6. C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In ICML, pages 1724-1732, 2017.
7. L. Lei, C.Ju, J.Chen, and M.I.Jordan. Non-convex finite-sum optimization via SCSG methods. In NIPS, pages 2345-2355, 2017.
8. Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods for deep learning. In IJCAI, pages 2955?2961, 2018.
9. Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In COLT, pages 1980-2022, 2017.