Escaping Saddle-points Faster under Interpolation

Abhishek Roy

MADDD Seminar, UC Davis

10.02.2020

Publication

[RBGM20] “Escaping Saddle-Point Faster under Interpolation-like Conditions”, Roy, Abhishek, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra. "Escaping Saddle-Points Faster under Interpolation-like Conditions." *arXiv preprint arXiv:2009.13016* (2020). (Accepted at *NeurIPS 2020*)

Nonconvex optimization in Statistics

- Nonconvex optimization is prevalent in statistics
- Finding global optima is difficult, even impossible in some cases
- In some cases, stationary points (local minima, saddle point, local maxima) have desirable statistical properties [Loh17; EG18; QCLP19]
- Examples: Piecewise affine regression, 2-layer neural network model with the ReLu activation function, robust regression, Smoothly Clipped Absolute Deviation (SCAD) penalty, Minimax Concave Penalty (MCP)
- There are methods to find stationary point, e.g., gradient descent

In recent past, lot of research on…

- ❑ Matrix Completion [GLM16, CL19]
- ❑ Robust Principle Component Analysis [GJZ17, FS19]
- ❑ Tensor Decomposition [GHJY15]
- ❑ Phase Retrieval [SQW16]
- ❑ Deep Neural Nets [NH17, Led20, LP20]

Nonconvex

Not all stationary points are desirable

Only local minima are statistically favorable !

Deep Learning and Nonconvex Optimization

Neural Tangent Kernel viewpoint

For ultra-wide network, randomly initialized gradient descent can be shown to work like kernel regression with the kernel NTK [JGH18]

All over-parametrized neural networks are not captured by NTK [ZLZ19]

NTK based results are for polynomially large networks (in depth and sample-size)

But, in practice, finite width work well as well

Landscape Analysis viewpoint

An alternative view for finite-width multilayer neural networks

All approximate local minima are also global minima [KK20, Led20, LP20]

Escape saddle points to reach local minima

Problem

$$\underset{\theta \in \mathbb{R}^d}{argmin} \left\{ f(\theta) \coloneqq \mathbf{E}_{\xi}[F(\theta, \xi)] \right\}$$

Nonconvex

Definition (ϵ-local minimum): Let $\lambda_{min}\left(\nabla^2 f(\theta)\right)$ is the minimum eigenvalue of $\nabla^2 f(\theta)$. A point $\bar{\theta}$ is called a **ϵ-local minimum/ ϵ-second-order stationary point** if,

$$\max\left(\sqrt{\left\|\nabla f(\bar{\theta})\right\|_2}, -\lambda_{min}\left(\nabla^2 f(\bar{\theta})\right)\right) \leq \sqrt{\epsilon}$$

Small Gradient
$\left(\left\|\nabla f(\bar{\theta})\right\|_2 \leq \epsilon\right)$

Locally Convex
$\left(\lambda_{min}\left(\nabla^2 f(\bar{\theta})\right) \geq -\sqrt{\epsilon}\right)$

Over-parametrized Models and Interpolation

Deep Neural Networks achieves zero training error

Empirical Loss $\mathrm{f}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(g_\theta(x_i) - y_i)^2$

g_θ: Parametrized space of functions

Over-parametrized Models:
$\#parameters \gg training\ data\ size$

Achieves perfect interpolation: $g_{\theta^*}(x_i) = y_i$

Interpolation [MBB18]: θ^* minimizes $\mathrm{f}(\theta) \Rightarrow \theta^*$ minimizes $(g_\theta(x_i) - y_i)^2$

Gradient Descent (GD): $\theta_{t+1} = \theta_t - \eta\nabla f(\theta_t)$

Minibatch Stochastic Gradient Descent (SGD):

$$\theta_{t+1} = \theta_t - \eta\nabla F(\theta_t, \xi)$$

Example: For empirical loss minimization $\frac{1}{m}\sum_{k=1}^{m}\nabla\left(g_{\theta_t}\left(x_{i_t}^k\right) - y_{i_t}^k\right)^2$, where $\{i_t\}_{k=1}^m$ are uniformly at random from $\{1,2,\cdots,n\}$.

Empirically Minibatch SGD performs better than GD

We theoretically show that, under interpolation, SGD does have faster convergence rate (even comparable with GD in some cases!) to reach the local minima of a non-convex loss function

First-order oracle: Outputs an estimate $\nabla F(\theta, \xi)$ of $\nabla f(\theta)$ such that $\mathbf{E}[\nabla F(\theta, \xi)] = \nabla f(\theta)$.

Strong Growth Condition (SGC) and main implication

For any point $\theta \in \mathbb{R}^d$, the stochastic gradient at θ, $\nabla F(\theta, \xi)$ satisfy

$$\boldsymbol{E_{\xi}\|\nabla F(\theta, \xi)\|_2^2 \leq \rho\|\nabla f(\theta)\|_2^2 \qquad (\rho > 1)} \quad \textbf{(SGC)} \qquad \text{[VBS18]}$$

With SGC, we have

$$\mathbf{E}\left[\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\nabla F(\theta_t, \xi_i) - \nabla f(\theta_t)\right\|_2^2\right] \leq \frac{\sigma^2}{n_1}$$

Previous Work on over-parametrized optimization

[SV09] Randomized version of the Kaczmarz method for consistent, overdetermined linear systems converges with exponential rate; similar condition like SGC

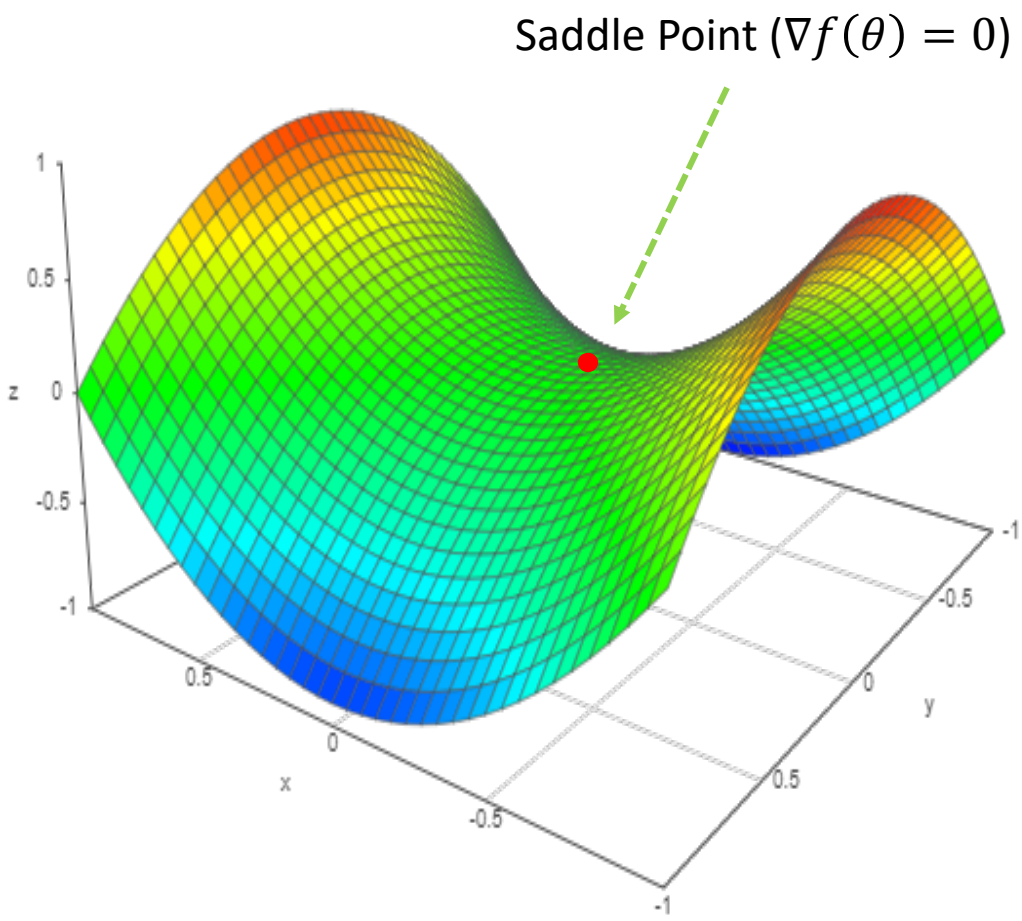
Interpolation regime:
[MBB18] showed mini-batch stochastic gradient descent (SGD) has exponential rates of convergence for unconstrained strongly-convex optimization, and linear rate for functions satisfying PL-inequality

SGC:
[MVL+20] Regularized subsampled Newton method (R-SSN) and the stochastic BFGS algorithm under SGC: global linear convergence for strongly convex case

[VBS18] constant step-size SGD has optimal convergence rate for strongly-convex and smooth convex functions

Non-convex setting: [VBS18] Constant step-size SGD can obtain the deterministic rate in the interpolation regime for converging to first-order stationary solution

How to escape from a saddle-point?



GD update:

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t)$$

Stuck at Saddle point!

Solution: Add random perturbation to the update

Perturbed Stochastic Gradient Descent (PSGD) Algorithm

Input: $\theta_0 \in \mathbb{R}^d$, η, r.
for $t = 0$ to T **do**
Set $g_t = \frac{1}{n_1} \sum_{i=1}^{n_1} g_{t,i}$ **where**

$$g_{t,i} = \nabla F\left(\theta_t, \xi_{t,i}\right) \qquad \text{(First-order)}$$

What if only noisy unbiased function-value estimates are available instead of noisy gradient ?

Sample $\beta_t \in \mathcal{N}\left(\mathbf{0}, r^2 \boldsymbol{I}_d\right)$
Update $x_{t+1} = x_t - \eta\left(g_t + \beta_t\right)$
end for

Perturbed Stochastic Gradient Descent (PSGD) Algorithm

Input: $\theta_0 \in \mathbb{R}^d$, η, r.
for $t = 0$ to T **do**
Set $g_t = \frac{1}{n_1} \sum_{i=1}^{n_1} g_{t,i}$ **where**

$$g_{t,i} = \nabla F\left(\theta_t, \xi_{t,i}\right) \quad \text{(First-order)}$$

$$g_{t,i} = \frac{F(\theta_t + \nu u_{t,i}, \xi_{t,i}) - F(\theta_t, \xi_{t,i})}{\nu} u_{t,i} \quad \text{(Zeroth-order)}$$

and $u_{t,i} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_d\right) \forall t = 1, 2, \cdots, T, i = 1, 2, \cdots, n_1$
Sample $\beta_t \in \mathcal{N}\left(\mathbf{0}, r^2 \boldsymbol{I}_d\right)$
Update $x_{t+1} = x_t - \eta\left(g_t + \beta_t\right)$
end for

Main Theorem: PSGD

Let f be a Lipschitz-continuous function with Lipschitz gradient and hessian. Let the noise in function-value/gradient is sub-gaussian. Then under SGC, for PSGD algorithm:

a) First-order: with probability at least $1-\delta$, total first-order oracle calls to reach ϵ-local minimizer are, $\tilde{O}(\epsilon^{-2})$.

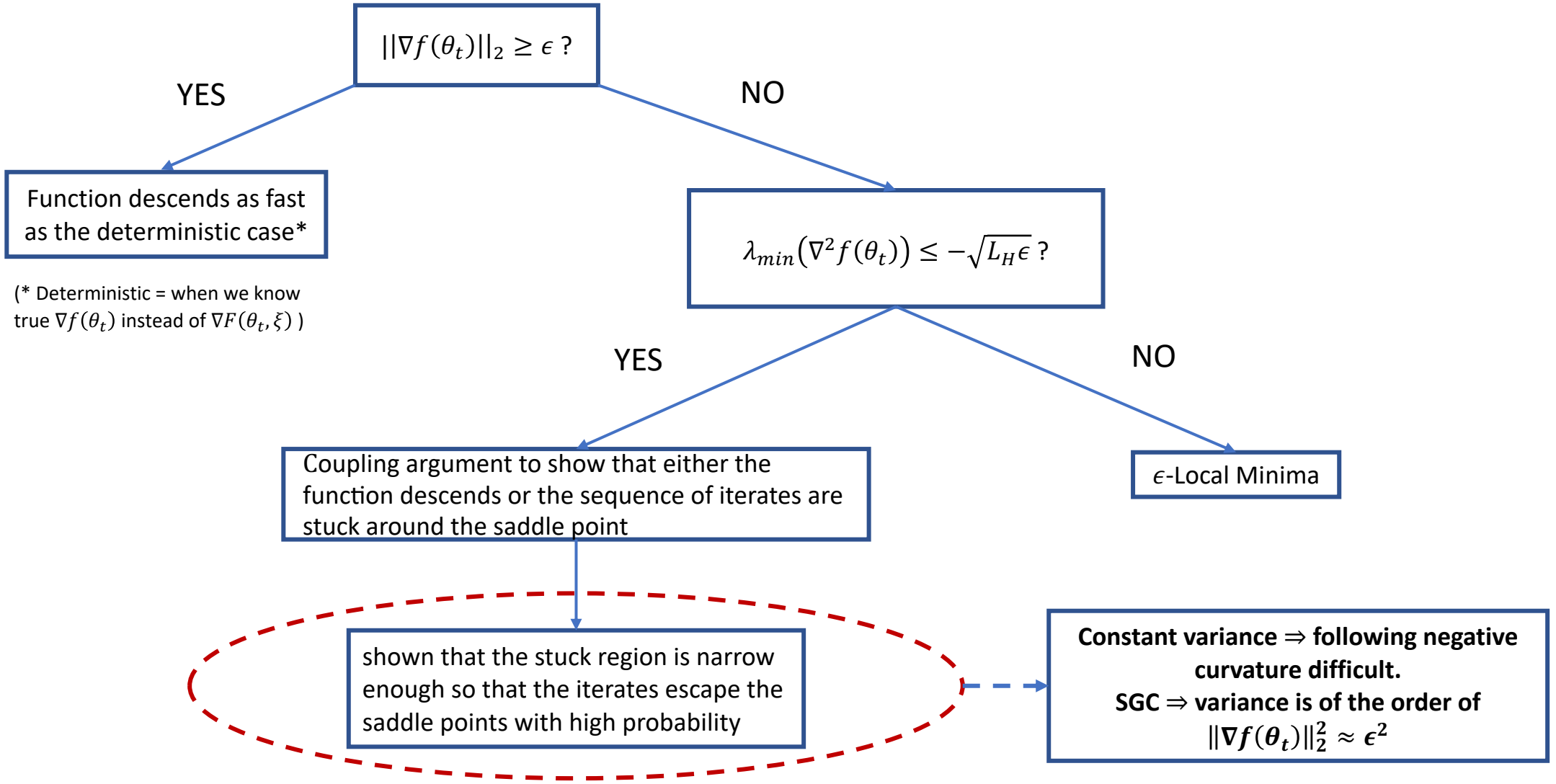
b) Zeroth-order: total corresponding zeroth-order calls are $\tilde{O}(d^{1.5}\epsilon^{-4.5})$.

c) Zeroth-order (Without SGC): total corresponding zeroth-order calls are $\tilde{O}(d^{1.5}\epsilon^{-5.5})$.

Algorithm	With SGC (This paper)		Without SGC		Deterministic
	ZO	HO	ZO	HO	HO
Perturbed GD	$\tilde{\mathcal{O}}(d^{1.5}\epsilon^{-4.5})$ This paper	$\tilde{\mathcal{O}}(\epsilon^{-2})$ This paper	$\tilde{\mathcal{O}}(d^{1.5}\epsilon^{-5.5})$ This paper	$\tilde{\mathcal{O}}(\epsilon^{-4})$ [JNG+19]	$\tilde{\mathcal{O}}(\epsilon^{-2})$ [JGN+17]

$$g_{t,i} = \frac{F(\theta_t + \nu u_{t,i}, \xi_{t,i}) - F(\theta_t, \xi_{t,i})}{\nu} u_i \rightarrow \alpha - sub\ exponential \longrightarrow$$ Worse concentration around $\nabla f(\theta)$

Proof Outline



So far…

- [MBB18] introduced the interpolation condition for over-parametrized models and analyzed SGD for convex functions
- [VBS18] introduced SGC and analyzed for non-convex functions but considers stationary points only
- Stationary points not enough in many applications but local-minima are
- We show faster convergence rate (matching deterministic) for PSGD to reach ϵ-local minima under SGC

Next…What happens for second-order methods?

Second-order Methods

- Second-order methods (uses hessian) have better convergence rate in the deterministic case
- Second-order oracle: Outputs an estimate $\nabla^2 F(\theta, \xi)$ such that $\mathbf{E}[\nabla^2 F(\theta, \xi)] = \nabla^2 f(\theta)$
- Cubic-regularized Newton Method.

Cubic-regularized Newton (CRN) Algorithm

Input: $\theta_1 \in \mathbb{R}^d, T, M, n_1, n_2$
for $t = 1$ to T **do**
Set $g_t = \frac{1}{n_1}\sum_{i=1}^{n_1} g_{t,i}$ **where**

$$g_{t,i} = \nabla F\left(\theta_t, \xi_{t,i}^G\right) \qquad \text{(First-order)}$$

$$g_{t,i} = \frac{F(\theta_t + \nu u_{t,i}^G, \xi_{t,i}^G) - F(\theta_t, \xi_{t,i}^G)}{\nu} u_i^G \qquad \text{(Zeroth-order)}$$

Set $H_t = \frac{1}{n_2}\sum_{i=1}^{n_2} H_{t,i}$ **where**

$$H_{t,i} = \nabla^2 F\left(\theta_t, \xi_{t,i}^H\right) \qquad \text{(Second-order)}$$

$$H_{t,i} = \frac{F(\theta_t + \nu u_{t,i}^H, \xi_{t,i}^H) + F(\theta_t - \nu u_{t,i}^H, \xi_{t,i}^H) - 2F(\theta_t, \xi_{t,i}^H)}{2\nu^2}\left(u_{t,i}^H u_{t,i}^{H\top} - I\right) \qquad \text{(Zeroth-order)}$$

where $u_{t,i}^{G[H]} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)\ \forall t = 1, 2, \cdots, T, i = 1, 2, \cdots, n_1[n_2]$
Update

$$\theta_{t+1} = \operatorname*{argmin}_z m_t\left(\theta_t, z, g_t, H_t, M\right),$$

where

$$m_t(z) = f(\theta_t) + (z - \theta_t)^\top g_t + \frac{1}{2}(z - \theta_t)^\top H_t (z - \theta_t) + \frac{M}{6}\|z - \theta_t\|^3$$

end for

Can be efficiently solved by Gradient Descent [CD16]

Main Theorem: CRN

Let f be a function with Lipschitz-gradient and hessian. Then under SGC, for CRN algorithm, we have:

a) Higher-order: choosing T, n_1, n_2, M appropriately, we get an ϵ-local minima

Total first-, and second-order oracle calls $\boldsymbol{O}\left(\frac{1}{\epsilon^{2.5}}\right)$.

b) Zeroth-order: choosing parameters appropriately, we get, total first–order oracle calls $\boldsymbol{O}\left(d\epsilon^{-2.5}\right)$, and second–order oracle calls $\boldsymbol{O}\left(d^4 \log d\, \epsilon^{-2.5}\right)$.

Algorithm	With SGC (This paper)		Without SGC		Deterministic
	ZO	FO+SO	ZO	HO	FO+SO
Perturbed GD	$\tilde{\mathcal{O}}\left(d^{1.5}\epsilon^{-4.5}\right)$ This paper	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$ This paper	$\tilde{\mathcal{O}}\left(d^{1.5}\epsilon^{-5.5}\right)$ This paper	$\tilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ [JNG+19]	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$ [JGN+17]
Cubic Newton	$\tilde{\mathcal{O}}\left(d^{4}\epsilon^{-2.5}\right)$ This paper	$\tilde{\mathcal{O}}\left(\epsilon^{-2.5}\right)$ This paper	$\tilde{\mathcal{O}}\left(d^{4}\epsilon^{-2.5}\right) + \mathcal{O}\left(d\epsilon^{-3.5}\right)$ [BG18]	$\tilde{\mathcal{O}}\left(\epsilon^{-3.5}\right)$ [TSJ+18]	$\mathcal{O}\left(\epsilon^{-1.5}\right)$ [NP06]

Table 1. Oracle complexities of PSGD and SCRN. ZO corresponds to number of calls to zeroth-order oracle and FO+SO corresponds to number of calls to first or second-order oracles. The result for PSGD and SCRN are given respectively in high-probability and in expectation.

Cubic-regularized Newton (CRN) Algorithm

Input: $\theta_1 \in \mathbb{R}^d, T, M, n_1, n_2$
for $t = 1$ to T **do**
Set $g_t = \frac{1}{n_1}\sum_{i=1}^{n_1} g_{t,i}$ **where**

$$g_{t,i} = \nabla F\left(\theta_t, \xi_{t,i}^G\right) \qquad \text{(First-order)}$$

$$g_{t,i} = \frac{F(\theta_t + \nu u_{t,i}^G, \xi_{t,i}^G) - F(\theta_t, \xi_{t,i}^G)}{\nu} u_i^G \qquad \text{(Zeroth-order)}$$

Set $H_t = \frac{1}{n_2}\sum_{i=1}^{n_2} H_{t,i}$ **where**

$$H_{t,i} = \nabla^2 F\left(\theta_t, \xi_{t,i}^H\right) \qquad \text{(Second-order)}$$

$$H_{t,i} = \frac{F(\theta_t + \nu u_{t,i}^H, \xi_{t,i}^H) + F(\theta_t - \nu u_{t,i}^H, \xi_{t,i}^H) - 2F(\theta_t, \xi_{t,i}^H)}{2\nu^2}\left(u_{t,i}^H u_{t,i}^{H\top} - I\right) \qquad \text{(Zeroth-order)}$$

→ No SGC-type condition here

where $u_{t,i}^{G[H]} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)\ \forall t = 1, 2, \cdots, T, i = 1, 2, \cdots, n_1[n_2]$
Update

$$\theta_{t+1} = \operatorname*{argmin}_z m_t\left(\theta_t, z, g_t, H_t, M\right),$$

where

$$m_t(z) = f(\theta_t) + (z - \theta_t)^\top g_t + \frac{1}{2}(z - \theta_t)^\top H_t (z - \theta_t) + \frac{M}{6}\|z - \theta_t\|^3$$

end for

References

[AZ17] Zeyuan Allen-Zhu, *Katyusha: The first direct acceleration of stochastic gradient methods*, The Journal of Machine Learning Research **18** (2017), no. 1, 8194–8244.

[AZL18] Zeyuan Allen-Zhu and Yuanzhi Li, *Neon2: Finding local minima via first-order oracles*, Advances in Neural Information Processing Systems, 2018, pp. 3716–3726.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song, *A convergence theory for deep learning via over-parameterization*, International Conference on Machine Learning, PMLR, 2019, pp. 242–252.

[BAA20] Qinbo Bai, Mridul Agarwal, and Vaneet Aggarwal, *Escaping saddle points for zeroth-order non-convex optimization using estimated gradient descent*, 2020 54th Annual Conference on Information Sciences and Systems (CISS), IEEE, 2020, pp. 1–6.

[BBM18] Raef Bassily, Mikhail Belkin, and Siyuan Ma, *On exponential convergence of sgd in non-convex over-parametrized learning*, arXiv preprint arXiv:1811.02564 (2018).

[BG18] Krishnakumar Balasubramanian and Saeed Ghadimi, *Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality and saddle-points*, arXiv preprint arXiv:1809.06474 (2018).

[CB18] Lenaic Chizat and Francis Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in neural information processing systems, 2018, pp. 3036–3046.

[CD16] Yair Carmon and John C Duchi, *Gradient descent efficiently finds the cubic-regularized non-convex newton step*, arXiv preprint arXiv:1612.00547 (2016).

[CGT11] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint, *Adaptive cubic regularisation methods for unconstrained optimization. part II: worst-case function-and derivative-evaluation complexity*, Mathematical programming **130** (2011), no. 2, 295–319.

[CL19] Ji Chen and Xiaodong Li, *Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel pca.*, Journal of Machine Learning Research **20** (2019), no. 142, 1–39.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach, *On lazy training in differentiable programming*, Advances in Neural Information Processing Systems, 2019, pp. 2937–2947.

[CRS17] Frank E Curtis, Daniel P Robinson, and Mohammadreza Samadi, *A trust region algorithm with a worst-case iteration complexity of $\epsilon^{-3/2}$ for nonconvex optimization*, Mathematical Programming **162** (2017), no. 1-2, 1–32.

[DB19] Aaron Defazio and Léon Bottou, *On the ineffectiveness of variance reduced optimization for deep learning*, Advances in Neural Information Processing Systems, 2019, pp. 1753–1763.

[DJL+17] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos, *Gradient descent can take exponential time to escape saddle points*, Advances in neural information processing systems, 2017, pp. 1067–1077.

[DLL+19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai, *Gradient descent finds global minima of deep neural networks*, International Conference on Machine Learning, 2019, pp. 1675–1685.

[DPG+14] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, Advances in neural information processing systems, 2014, pp. 2933–2941.

[FLLZ18] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang, *Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator*, Advances in Neural Information Processing Systems, 2018, pp. 689–699.

[FLZ19] Cong Fang, Zhouchen Lin, and Tong Zhang, *Sharp analysis for nonconvex sgd escaping from saddle points*, Conference on Learning Theory, 2019, pp. 1192–1234.

[FS19] Salar Fattahi and Somayeh Sojoudi, *Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis*, 2019.

[FVGP19] Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras, *Efficiently avoiding saddle points with zero order methods: No gradients required*, arXiv preprint arXiv:1910.13021 (2019).

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan, *Escaping from saddle points: online stochastic gradient for tensor decomposition*, Conference on Learning Theory, 2015, pp. 797–842.

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, arXiv preprint arXiv:1704.00708 (2017).

[GL13] Saeed Ghadimi and Guanghui Lan, *Stochastic first-and zeroth-order methods for nonconvex stochastic programming*, SIAM Journal on Optimization **23** (2013), no. 4, 2341–2368.

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma, *Matrix completion has no spurious local minimum*, Advances in Neural Information Processing Systems, 2016, pp. 2973–2981.

[HV15] Benjamin D Haeffele and René Vidal, *Global optimality in tensor factorization, deep learning, and beyond*, arXiv preprint arXiv:1506.07540 (2015).

[HYV14] Benjamin Haeffele, Eric Young, and Rene Vidal, *Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing*, International conference on machine learning, 2014, pp. 2007–2015.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, Advances in neural information processing systems, 2018, pp. 8571–8580.

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan, *How to escape saddle points efficiently*, Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1724–1732.

[JNG+19] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan, *On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points*, arXiv preprint arXiv:1902.04811 (2019).

[JNJ18] Chi Jin, Praneeth Netrapalli, and Michael I Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, Conference On Learning Theory, 2018, pp. 1042–1085.

[Kaw16] Kenji Kawaguchi, *Deep learning without poor local minima*, Advances in neural information processing systems, 2016, pp. 586–594.

[KK20] Kenji Kawaguchi and Leslie Kaelbling, *Elimination of all bad local minima in deep learning*, International Conference on Artificial Intelligence and Statistics, 2020, pp. 853–863.

[L+17] Po-Ling Loh et al., *Statistical consistency and asymptotic normality for high-dimensional robust m-estimators*, The Annals of Statistics **45** (2017), no. 2, 866–896.

[Led20] Johannes Lederer, *No spurious local minima: on the optimization landscapes of wide and deep neural networks*, 2020.

[LP20] Jonathan Lacotte and Mert Pilanci, *All local minima are global for two-layer relu neural networks: The hidden convex optimization landscape*, 2020.

[LPP+17] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht, *First-order methods almost always avoid saddle points*, arXiv preprint arXiv:1710.07406 (2017).

[LRY+19] Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, and Mingyi Hong, *Snap: Finding approximate second-order stationary solutions efficiently for non-convex linearly constrained problems*, arXiv preprint arXiv:1907.04450 (2019).

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht, *Gradient descent only converges to minimizers*, Conference on learning theory, 2016, pp. 1246–1257.

[LZHH19] Songtao Lu, Ziping Zhao, Kejun Huang, and Mingyi Hong, *Perturbed projected gradient descent converges to approximate second-order points for bound constrained nonconvex problems*, ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5356–5360.

[MBB18] Siyuan Ma, Raef Bassily, and Mikhail Belkin, *The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning*, International Conference on Machine Learning, 2018, pp. 3325–3334.

[MOJ18] Aryan Mokhtari, Asuman Ozdaglar, and Ali Jadbabaie, *Escaping saddle points in constrained optimization*, Advances in Neural Information Processing Systems, 2018, pp. 3629–3639.

[MVL+20] Si Yi Meng, Sharan Vaswani, Issam Laradji, Mark Schmidt, and Simon Lacoste-Julien, *Fast and furious convergence: Stochastic second order methods under interpolation*, arXiv preprint arXiv:1910.04920 (2020).

[NH17] Quynh Nguyen and Matthias Hein, *The loss surface of deep and wide neural networks*, arXiv preprint arXiv:1704.08045 (2017).

[NP06] Yurii Nesterov and Boris T Polyak, *Cubic regularization of newton method and its global performance*, Mathematical Programming **108** (2006), no. 1, 177–205.

[NR19] Maher Nouiehed and Meisam Razaviyayn, *A trust region method for finding second-order stationarity in linearly constrained non-convex optimization*, arXiv preprint arXiv:1904.06784 (2019).

[NS17] Yurii Nesterov and Vladimir Spokoiny, *Random gradient-free minimization of convex functions*, Foundations of Computational Mathematics **17** (2017), no. 2, 527–566.

[NWS14] Deanna Needell, Rachel Ward, and Nati Srebro, *Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm*, Advances in neural information processing systems, 2014, pp. 1017–1025.

[Pol63] Boris Teodorovich Polyak, *Gradient methods for minimizing functionals*, Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki **3** (1963), no. 4, 643–653.

[QCLP19] Zhengling Qi, Ying Cui, Yufeng Liu, and Jong-Shi Pang, *Statistical analysis of stationary solutions of coupled nonconvex nonsmooth empirical risk minimization*, arXiv preprint arXiv:1910.02488 (2019).

[RBGM19] Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra, *Multi-point bandit algorithms for nonstationary online nonconvex optimization*, arXiv preprint arXiv:1907.13616 (2019).

[SBG19] Lingqing Shen, Krishnakumar Balasubramanian, and Saeed Ghadimi, *Non-asymptotic results for langevin monte carlo: Coordinate-wise and black-box sampling*, arXiv preprint arXiv:1902.01373 (2019).

[Sch20] Mark Schmidt, *Faster algorithms for deep learning? (presentation in vector institute: https://www.cs.ubc.ca/ schmidtm/documents/2020_ vector_ smallresidual.pdf)*, 2020.

[SQW18] Ju Sun, Qing Qu, and John Wright, *A geometric analysis of phase retrieval*, Foundations of Computational Mathematics **18** (2018), no. 5, 1131–1198.

[Sun19] Ruoyu Sun, *Optimization for deep learning: theory and algorithms*, arXiv preprint arXiv:1912.08957 (2019).

[SV09] Thomas Strohmer and Roman Vershynin, *A randomized kaczmarz algorithm with exponential convergence*, Journal of Fourier Analysis and Applications **15** (2009), no. 2, 262.

[TSJ+18] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan, *Stochastic cubic regularization for fast nonconvex optimization*, Advances in neural information processing systems, 2018, pp. 2899–2908.

[VBS18] Sharan Vaswani, Francis Bach, and Mark Schmidt, *Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron*, arXiv preprint arXiv:1810.07288 (2018).

[WZLL18] Zhe Wang, Yi Zhou, Yingbin Liang, and Guanghui Lan, *Stochastic variance-reduced cubic regularization for nonconvex optimization*, arXiv preprint arXiv:1802.07372 (2018).

[ZCZG20] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu, *Gradient descent optimizes over-parameterized deep relu networks*, Machine Learning **109** (2020), no. 3, 467–492.

[ZG19] Dongruo Zhou and Quanquan Gu, *Stochastic recursive variance-reduced cubic regularization methods*, arXiv preprint arXiv:1901.11518 (2019).

[ZXZ18] Junyu Zhang, Lin Xiao, and Shuzhong Zhang, *Adaptive stochastic variance reduction for subsampled newton method with cubic regularization*, arXiv preprint arXiv:1811.11637 (2018).

Thank You