ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 5-2: Star-Convexity and α-Weak-Quasi-Convexity

Jia (Kevin) Liu

Assistant Professor
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

Spring 2022

Outline

In this lecture:

$O\left(\frac{1}{K^2}\right)$.

- Star-Convexity and α-Weak-Quasi-Convexity
- Convergence Results under Star-Convexity and α-Weak-Quasi-Convexity

Star-Convex Function

Definition 1 ([Nesterov and Polyak, Math Prog'06])

A function $f(\mathbf{x})$ is called star-convex if for some global minimizer $\mathbf{x}^*$ and for all $\lambda \in [0,1]$ and $\mathbf{x} \in \mathbb{R}^d$

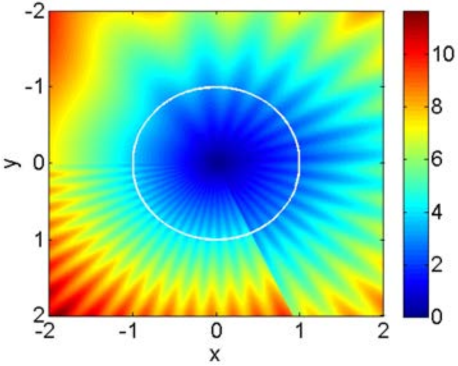
$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{x}^*) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{x}^*)$$

Remarks

- Any interval connecting some point $\mathbf{x}$ and some global minimizer $\mathbf{x}^*$ lies not lower than the graph
- Considerably weaker than convexity
- For example, $|x|(1-e^{-|x|})$ is a nonconvex star-convex function.

An Example of Star-Convex Landscape

- Intuitively, if we visualize the objective function as a landscape, star-convexity means that the global optimum is "visible" from every point (i.e., "no blocking ridges", figure from [Lee and Valiant, FOCS'16])



Optimal First-Order Algorithms under Star-Convexity

- AGMsDR [Nesterov et al., arXiv:1809.05895]
  - Accelerated Gradient Method with Small-Dimensional Relaxation (AGMsDR)
  - For star-convex L-smooth functions, AGMsDR achieves

$$\min_{k=\lceil T/2\rceil,\ldots,T} \|\nabla f(\mathbf{y}_k)\|_*^2 \leq \frac{64L^2\Delta_0}{T^3},$$

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{4L\Delta_0}{T^2} = O\left(\frac{1}{T^2}\right)$$

α-Weak-Quasi-Convex Function

A more general class of functions:

Definition 2

A function $f(\mathbf{x})$ is called α-weakly-quasi-convex function if for some global minimizer $\mathbf{x}^*$, some some $\alpha \in (0, 1]$, and $\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x})$ satisfies

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{\alpha}\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle$$

Remarks

- Continuously differentiable 1-weakly-quasi-convex functions are exactly the star-convex functions [Guminov et al., arXiv:1710.00797]

Optimal FO Methods under α-Weak-Quasi-Convexity

Iteration complexity: $O(\frac{1}{T^2})$

- AGMsDR [Nesterov et al., arXiv:1809.05895]: $O(\alpha^{-3/2}L^{1/2}\Delta_0\epsilon^{-1/2})$
  - AGMsDR requires exact line search
- SESOP [Guminov et al., arXiv:1710.00797]: $O(\alpha^{-1}L^{1/2}\Delta_0\epsilon^{-1/2})$
  - SESOP requires exact line search
- GAGD [Hinder et al., COLT'20]: $O(\alpha^{-1}L^{1/2}\Delta_0\epsilon^{-1/2})$
  - GAGD only requires simple backtracking and binary line search
  - Also provided iteration complexity lower bound, thus proving GAGD being order-optimal in terms of iteration complexity

(α, μ)-Strongly Quasi-Convex Function

A more general class of functions: linear

Definition 3

A function $f(\mathbf{x})$ is called α-weakly-quasi-convex function if for some global minimizer $\mathbf{x}^*$, some some $\alpha \in (0, 1]$, and $\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x})$ satisfies

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{\alpha}\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$$

Iteration complexity:

- GAGD [Hinder et al., COLT'20]: $O(\alpha^{-1} L^{1/2} \Delta_0 \log(\alpha^{-1}\epsilon^{-1}))$

Stochastic Methods under α-Weak-Quasi-Convexity

- SGD [Gower et al., AISTATS'21]: finite-sum minimization:
  - Sample complexity bound under "expected residual" assumption: $\mathbb{E}[\|g(\mathbf{x}) - g(\mathbf{x}^*) - (\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*))\|^2] \le 2\rho(f(\mathbf{x}) - f(\mathbf{x}^*))$ for some $\rho > 0$:

$$O\left(\frac{(\rho + L)\Delta_0^2}{\alpha^2 \epsilon} + \frac{\sigma_*^2 \Delta_0^2}{\alpha^2 \epsilon^2}\right)$$

  - Under interpolation condition and with Polyak step-size:

$$O\left(\frac{\bar{L}\Delta_0^2}{\alpha^2 \epsilon}\right)$$

    - ⋆ $\bar{L}$ is the expected smoothness constant
    - ⋆ In full batch case (i.e., $g(\mathbf{x}) = \nabla f(\mathbf{x})$), we have $\bar{L} = L$
    - ⋆ in importance sampling case (i.e., $g(\mathbf{x}) = \nabla f_j(\mathbf{x})$ where $j = i$ with prob. $L_i / \sum_{k=1}^N L_k$), we have $\bar{L} = \frac{1}{N}\sum_{i=1}^N L_i$

Thank You!