Single Loop Gaussian Homotopy for Non-convex Functions

Hidenori Iwakiri

Joint work with Yuhang Wang, Shinji Ito, Akiko Takeda

Introduction

Problem setting

- f : nonconvex function

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x)$$

Gaussian homotopy (GH)

- Method to find better stationary points for non-convex optimization using Gaussian smoothing

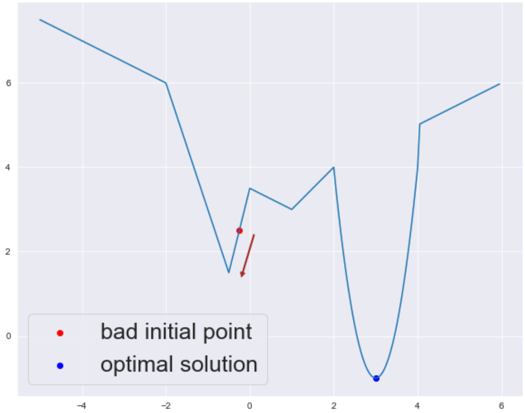
Gaussian smoothing

- $u \sim \mathcal{N}(0, \mathrm{I}_d)$
- $t > 0$: smoothing parameter (larger -> smoother)

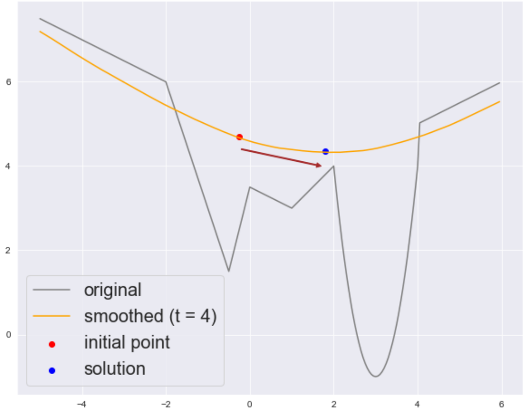
$$F(x, t) := E_u[f(x + tu)]$$

Toy example to understand Gaussian homotopy

Problem: GD based method cannot reach optimal solution
when starting from a bad initial point

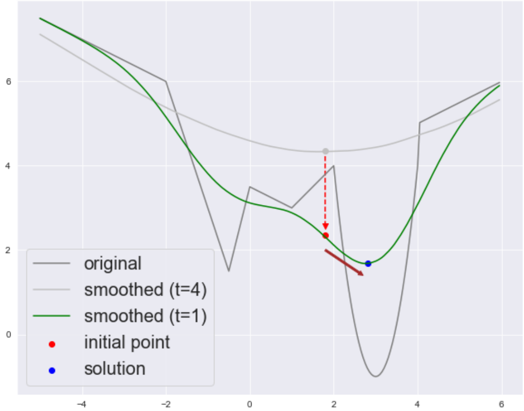


Toy example to understand Gaussian homotopy



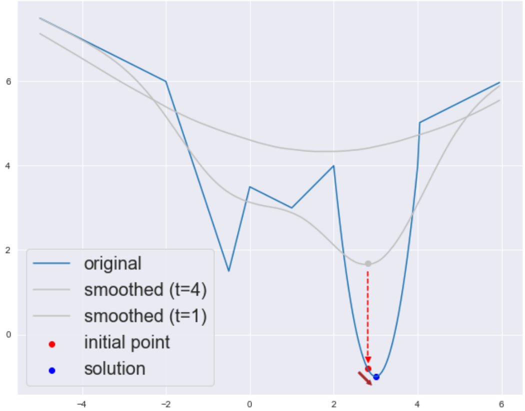
Optimize a simpler smoothed function

Toy example to understand Gaussian homotopy



Decrease the smoothing parameter t and optimize a function closer to the original one from the previous solution

Toy example to understand Gaussian homotopy



By repeating the similar procedure, the algorithm has successfully found the optimal solution!

Problems of previous work

- There exists some works [Chen, 2012; Hazan et al., 2016; Mobahi et al., 2015] that give theoretical analyses of Gaussian homotopy.
- However, they have not analyzed the convergence rate or the function class to be analyzed is limited
- Moreover, all of them consider double loop approach, which requires high computational costs

Algorithm 1 Double loop Gaussian homotopy method

```
Require: Iteration number K, initial point x_0,
         sequence {t_1, ..., t_K} satisfying t_1 > ... > t_K.
// Outer loop
for k = 1, ...K do
    // Inner loop
    Find a stationary point x_k of F(x, t_k)
     with the initial solution x_{k-1}.
return x_K
```

Contributions

- Propose novel single loop GH (SLGH) algorithms and analyze their convergence rates to an ϵ-stationary point
  - SLGH algorithms become faster than a double loop one by around its number of outer loops.
  - This is the first analysis of convergence rates of GH methods for general non-convex problems
- Propose zeroth-order SLGH (ZOSLGH) algorithms based on zeroth-order estimators of gradient and Hessian values
  - Useful when calculation of Gaussian smoothing is difficult
- Check the performance of SLGH on numerical experiments (artificial non-convex examples, black-box adversarial attacks)
  - Converges much faster than an existing double loop GH
  - Able to find better solutions than GD-based methods.

Proposed single loop algorithms (first-order)

Algorithm 2 SLGH (Single Loop Gaussian Homotopy)

Choose initial solution x_0 and initial smoothing parameter t_0.
for $k = 1, \ldots K$ **do**
    Query a gradient oracle $G_x = \nabla_x F(x_{k-1}, t_{k-1})$
    Query a derivative oracle $G_t = \frac{\partial F(x_{k-1}, t_{k-1})}{\partial t}$
    Update x_k by

$$x_k = x_{k-1} - \beta_k G_x$$

    Update t_k by

$$t_k = \begin{cases} \max\{0, \min\{t_{k-1} - \eta_k G_t,\ \gamma t_k\}\} & (\mathrm{SLGH_d}) \\ \gamma t_k & (\mathrm{SLGH_r}) \end{cases}$$

return $\hat{x} = x_{k'},\ k' = \mathrm{argmin}_{k \in \{0, \ldots, K\}} \|\nabla f(x_k)\|^2$

Theoretical analysis (first-order)

Theorem (Convergence analysis for SLGH)

Suppose Assumption A1 holds, and let $\hat{x} := x_{k'}$, $k' = \operatorname{argmin}_{k \in [T]} \|\nabla f(x_k)\|$. *Set the stepsize for* x *as* $\beta = 1/L_1$ $(L_1$: *smoothness parameter of* $f)$.

Then, for any setting of the parameter γ*, the output* $\hat{x}$ *satisfies* $\|\nabla f(\hat{x})\| \leq \epsilon$ *with the iteration complexity of*

$$T = O\left(d^{3/2}/\epsilon^2\right).$$

Further, if we choose $\gamma \leq d^{-\Omega(\epsilon^2)}$*, the iteration complexity can be bounded as*

$$T = O(1/\epsilon^2) \text{ (= iteration complexity of GD)}.$$

Experiment: adversarial attack

Black-box adversarial attack problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \ell(0.5\tanh(\tanh^{-1}(2a)+x) \\ + \lambda\|0.5\tanh(\tanh^{-1}(2a)+x)-a\|^2$$

- $a \in \mathbb{R}^d$: input image
- $\ell : \mathbb{R}^d \to \mathbb{R}$: attack loss
- $\lambda > 0$: regularization hyperparameter
- x: noise

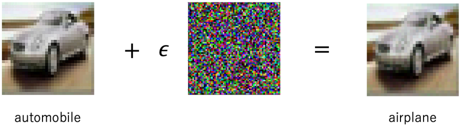


Figure: Adversarial attack example

Results (Dataset: CIFAR-10, $N = 100$)

- Initial point $x_0 : 0$ (no-noise, local minimum)

	succ rate	iters to 1st succ	total loss
ZOSGD	0.88	835	27.70
ZOAdaMM	0.85	3335	20.24
ZOGradOpt	0.65	6789	41.45
ZOSLGH$_r$ ($\gamma = 0.999$)	<u>0.93</u>	4979	14.26
ZOSLGH$_d$ ($\gamma = 0.999, \eta = 1e^{-4}$)	<u>0.92</u>	4436	16.49

- Single loop GHs achieve higher succ rates than SGD algos
  - Can escape the local minima ($x = 0$) due to sufficient smoothing
- Single loop GHs achieve higher succ rates and fewer iters to 1st success than double loop GH
  - Single loop structure requires lower computational costs

References I

- Chen, X. (2012):
  Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical programming*, *134*(1), 71–99.
- Hazan, E., Levy, K. Y., & Shalev-Shwartz, S. (2016):
  On graduated optimization for stochastic non-convex problems. *International conference on machine learning*, *48*, 1833–1841.
- Mobahi, H., & Fisher, J. W. (2015):
  On the link between gaussian homotopy continuation and convex envelopes. *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 43–56.