Research of global optimization in machine Learning

기계 학습의 전역 수렴 분석 방법론

Analysis Framework of global convergence in machine learning

- Next-Generation System SW Research Section,
  Future Computing Research Division,
  Artificial Intelligence Research Laboratory.
  Electronics and Telecommunications Research Institute (ETRI)
  Daejeon, Korea
  jnwseok@etri.re.kr

Jinwuk Seok

2022.8.02

Introduction

Why should we analyze a global property in machine learning?

- Requirement of an algorithm with faster and better optimization performance
    - Requirement of a robust algorithm to **Local minima**
    - **Robust algorithm to any initial points**

Is Global Optimization Possible in Machine Learning?

- Difficulty in Analysing method
    - **Limitation of analysis on the Hilbert Space**
    - It requires the analysis on a **normed (or measure-based) space**
        - Requires heavily **complicated analysis**
        - Limitation of **compactification method** for an objective function
        - Entropy based objective function, **KL-divergence**
    - Limitations of Heuristic methods
    - Complication of Probability methods

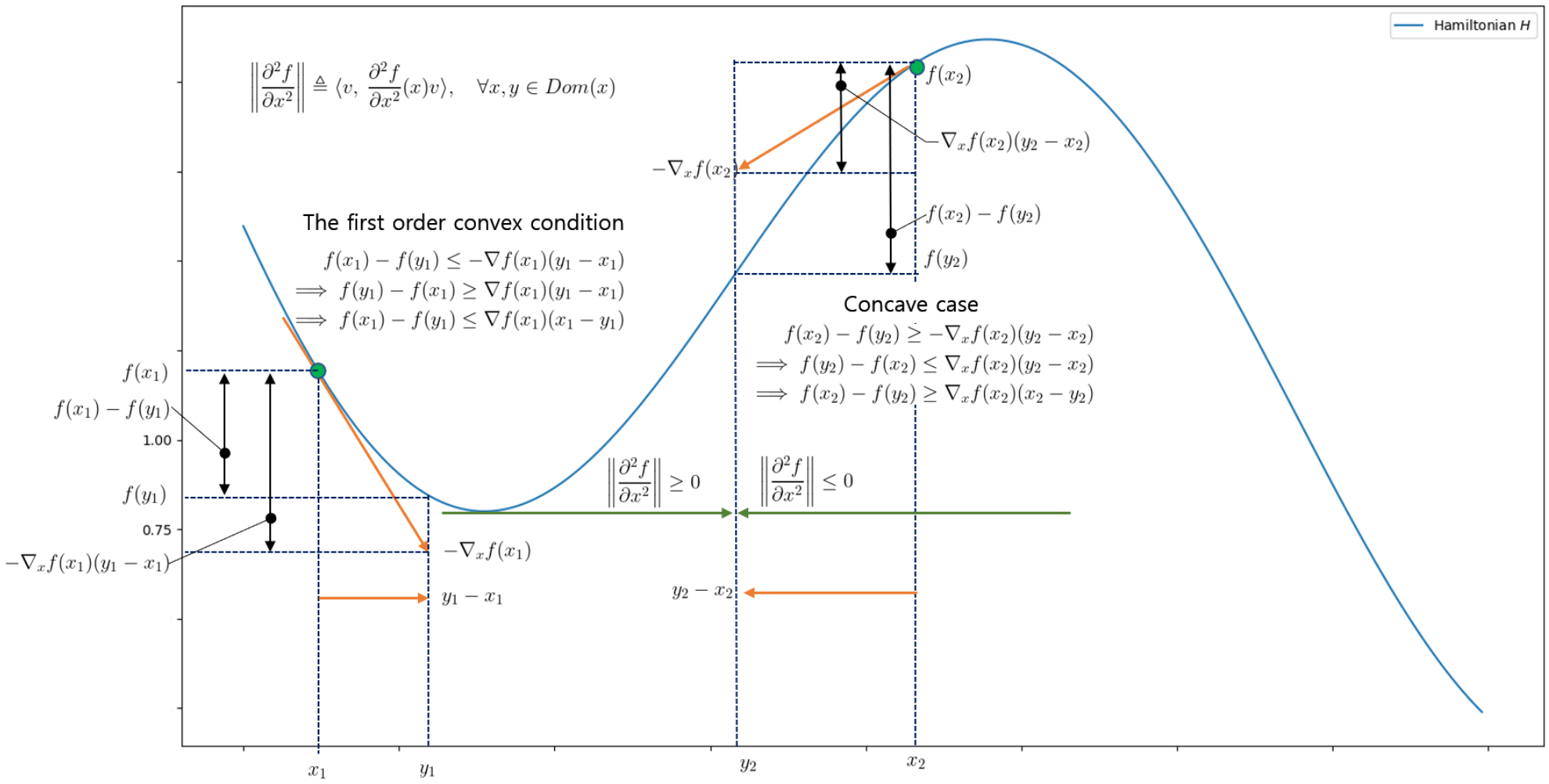
Known Global Optimization Algorithms

- Markov Chain Monte Carlo (MCMC)
  - Simulated Annealing
- Heuristic Algorithms
  - Evolutionary (Genetic) Algorithm, Particle Swarm Algorithms
- Quantum Computing (Qunatum Annealing)
- Baysian Optimization
- Quantization Based Optimization [3]

What is the mathematical tools for global analysis.

- Topological Space used in the conventional analysis is not appropriate.
  - Analysis on **Hilbert space** (is not sufficient)
    - Inner product is main tools for analysis
    - **Convex properties** is not available.
  - **Lipschitz Continuous** is most important assumptions
    - Holder Continuous is Alternative tools
    - Pathewise continuous is another tools
    - Sobolev space and Lebesgue measurable space
    - At worst, we should analyse on **the Banach space**.

Problems in the conventional analysis



Convex Assumption

- Assume that

$$\exists m, M \in \mathbf{R}^{+} \quad \text{such that} \quad m \leq \left\| \frac{\partial^2 f}{\partial x^2} \right\| < M$$

, where $\left\| \frac{\partial^2 f}{\partial x^2} \right\|$ is defined as

$$\forall v \in \mathbf{R}^{n} \ \text{ such that } \ \|v\| = 1, \quad \left\| \frac{\partial^2 f}{\partial x^2} \right\| \triangleq \langle v, \frac{\partial^2 f}{\partial x^2} v \rangle.$$

$\forall x_t \in \mathbf{R}^{n}$, we set the learnig equation as follows :

$$x_{t+1} = x_t + \varepsilon h(x_t)$$

with the directional derivative $h(x_t) = -\nabla f(x_t)$, and the learnig rate $\varepsilon \in \mathbf{R}(0, 1)$.

$$\begin{aligned} f(x_{t+1}) - f(x_t) &= \langle \nabla f(x_t), \varepsilon h(x_t) \rangle + \int_0^1 (1-s) \langle \varepsilon h(x_t), \frac{\partial^2 f}{\partial x^2} \varepsilon h(x_t) \rangle \\ &\leq -\varepsilon \|\nabla f(x_t)\|^2 + \frac{M}{2} \varepsilon^2 \|\nabla f(x_t)\|^2 ds \\ &= \frac{M}{2} \|\nabla f(x_t)\|^2 \varepsilon \left(\varepsilon - \frac{2}{M} \right) < 0, \quad \therefore 0 < \varepsilon < \frac{2}{M} \end{aligned}$$

- For the concave case such that the eigen-value of the Hessian of the objective function $f(x)$ is negative,
    - The equation (4) is not no more hold.
    - In other words, $\exists m, M \in \mathbf{R}^{+} \quad \text{such that} \quad -M \leq \left\| \frac{\partial^2 f}{\partial x^2} \right\| < -m$

$$
\begin{aligned}
f(x_{t+1}) - f(x_t) &= \langle \nabla f(x_t), \varepsilon h(x_t) \rangle + \int_0^1 (1-s) \langle \varepsilon h(x_t), \frac{\partial^2 f}{\partial x^2} \varepsilon h(x_t) \rangle ds \\
&\leq -\varepsilon \|\nabla f(x_t)\|^2 - \frac{m}{2} \varepsilon^2 \|\nabla f(x_t)\|^2 \\
&= -\frac{m}{2} \|\nabla f(x_t)\|^2 \varepsilon \left(\varepsilon + \frac{2}{m} \right) < 0, \qquad \therefore \varepsilon < -\frac{2}{m} \text{ or } \varepsilon > 0
\end{aligned}
$$

    - The above analysis shows the analysis result for the input on the concave domain is the equal to the convex case.
- Most of all, **the analysis under the convex assumption cannot provide any information of global optimization.**
    - **Limitation of the analysis on the Hilbert Space**

Approaching the Analysis of the Global Optimization

Analysis on a measurable space

- Analysis of a Hilbert space under the convex assumption is a comparison on a scalar field.
    - In other words, it is the comparison with norms.
- Through the analysis on a normed space or measurable space, we can overcome the limitation of analysis on a conventional Hilbert space.
- Representative Measurable Space
    - **Banach Space**
    - **Probability Space**
        - By ergodicity, the analysis of time-invariant big data and the analysis of a stochastic process can be equal.
        - Compared to other analysis methodology, there is a lot of well-known mathematical analysis technique.

Limitaion of Hilbert Space based Analysis : for Instance

- Consider the following learning equation amended with other momentum or reinforcement term.

$$x_{t+1} = x_t + \varepsilon h(x_t) + \lambda g(t)$$

- Let the following assumptions for convenice in anaysis.
  - There exists a positive value α such that $\alpha\|g(t)\| < \|h(x_t)\|,\ \forall x_t \in \mathbf{R}^n, \text{and } t \in \mathbf{R}^+$.
  - There exists a positive value η such that $\varepsilon > \eta\lambda$.
  - The learning rate $\varepsilon \in \mathbf{R}(0, \frac{1}{M})$
- For the above learning equation, there exists a hill-climbing effect caused by the condition $o(M) < -\left(\varepsilon - \frac{1}{M}\right)^2$ such that $f(x_{t+1}) - f(x_t) > 0$ under the convex assumption.

$$f(x_{t+1}) - f(x_t) < \frac{M}{2}\|\nabla f(x_t)\|^2 \left(\left(\varepsilon - \frac{1}{M}\right)^2 + o(M)\right)$$

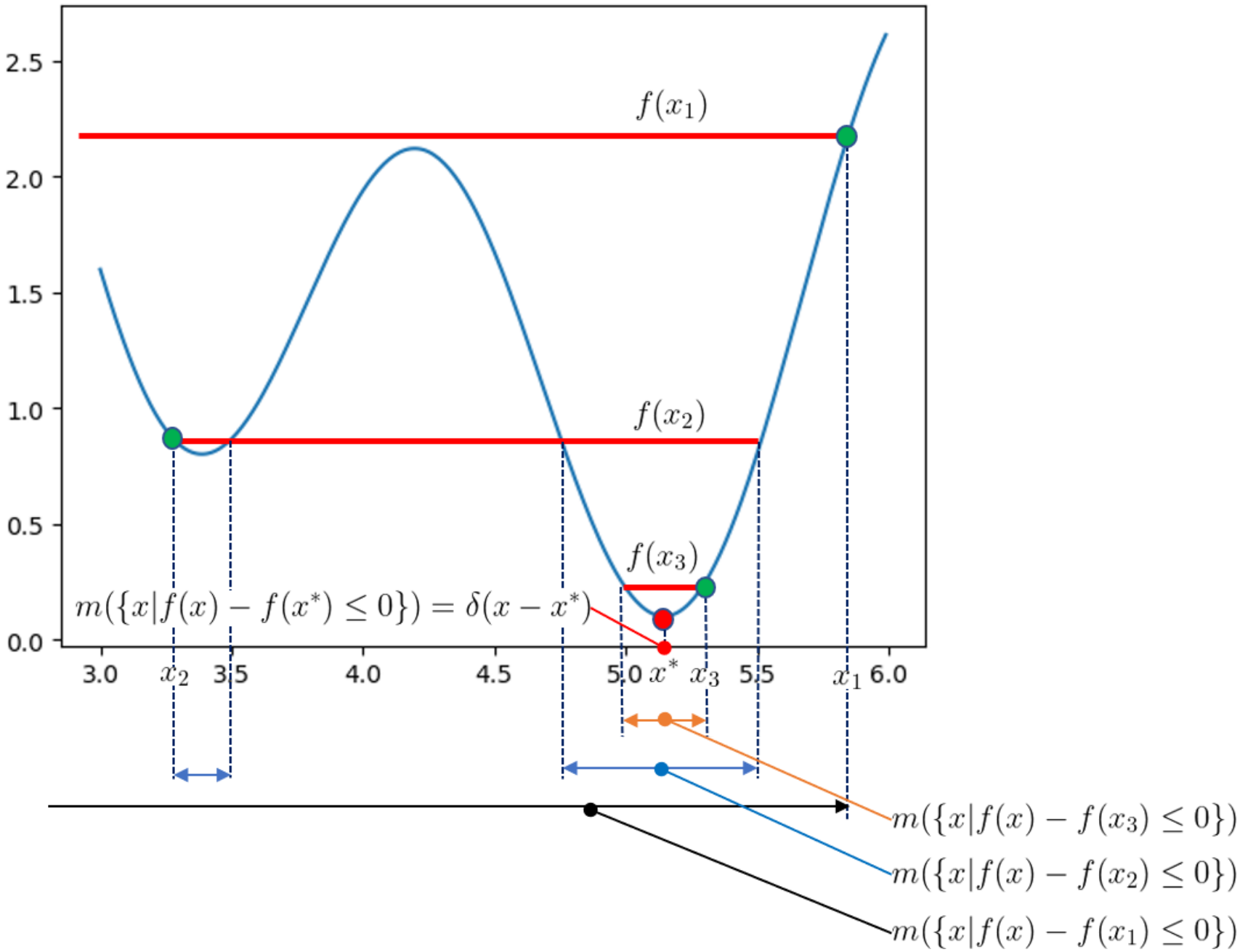
$$\because \exists M > 0, \text{ such that } o(M) < \frac{\lambda^2}{\alpha^2} \cdot C_0$$

- For a Concave case, if $o(m) > \left(\varepsilon + \frac{1}{m}\right)^2$, there exists the hill-climbing such that $f(x_{t+1}) - f(x_t) > 0$, as follows:

$$f(x_{t+1}) - f(x_t) < -\frac{m}{2}\|\nabla f(x_t)\|^2 \left(\left(\varepsilon + \frac{1}{m}\right)^2 - o(m^2)\right)$$
$$\because \exists m > 0, \text{ such that } o(m) > 0$$

- However, both analyses only provide **the existence of hill-climbing**.
- **The above analysis cannot provide any information on the global property**.
    - Since the function $o(\cdot)$ can be varied by a learning equation, it would be possible to analyze the global properties in more detail.

Concept of Analysis on Measurable Space : Weak Convergence



- Let $D_k = \{x | f(x) - f(x_k) \leq 0\}$

$$\Omega \supset D_1 \supset D_2 \cdots \supset D_\infty \supseteq X^*,$$

  - Let a (Lebesgue) measure m for D_k. Then, we obtain

$$1 = m(\Omega) > m(D_1) > m(D_2) \cdots > m(D_\infty) = \delta(x - x^*)$$

- The measure of the domain to a specific level set should be converged to the $\delta(x)$ function, as presented in the figure.
- In addition, **prove the equality of the measure to the level set at different points**.

$$\forall x, y \in D(x^*),\ x \neq y$$
$$|m(\{y | f(y) - f(x^*) \leq 0\}) - m(\{x | f(x) - f(x^*) \leq 0\})| = 0$$

- In other words, if we prove the **Weak Convergence** of an algorithm, it provides the **Global Convergence** of the algorithm.

Practical Perspective : Ananlysis on Probability Space

- To hold generality, we introduced a Lebesgue measure previously.
- However, to analyze a practical algorithm, we employ the **probability measure**.
    - Practically, since an algorithm processes data sequentially, we can regard the processed data as a stochastic process with ergodicity.
    - Therefore, we can do an analysis on filtration $\mathcal{F}_t,\ t \geq 0$ on a σ-Algebra $\mathcal{F}$.
    - Additionally, the topology for analysis is **Probability Space** $(\Omega, \mathcal{F}, P)$. (or $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$)
- Merit of analysis on probability space.
    - We can use many mathematical tools.
        - Many Probability Inequalities (such as Holder inequality, Markov inequality, and so on.)
        - Stochastic and Geometric Calculus
    - We construct **machine learning** or **artificial intelligence** based on neural networks on **probability assumptions**.

Laplace's Method

Let Q be a fixed probability measure on $(\Omega, \mathcal{F})$, and f is a continuous function on a compact set $\Omega \subset \mathbf{R}^n$ with the following assumption :

$$Q\{x|f(x) - a < 0\}, \text{ if } a > \inf_x f(x).$$

- A set N is related to f by $N = \{x|f(x) = \inf_y f(y)\}$.
- When there exists a probability P on N , **Laplace's method** is interpreted as **weak convergence of probability measure** [1].

$$\lim_{\tau \to \infty} \sup_{x_t, x_{t+\tau} \in \mathbf{R}^n} |p(t, \bar{x}_t, t+\tau, x^*) - p(t, x_t, t+\tau, x^*)| = 0$$

- For an instance of Laplace's method, suppose that a Radon-Nycodym derivative is given as follows

$$m^{\theta}(x)dx \triangleq \frac{dP_{\theta}}{dQ}(x) = \exp\left(-\frac{f(x)}{\theta}\right)\left(\int_{\Omega}\exp\left(-\frac{f(x)}{\theta}\right)dQ(x)\right)^{-1}$$

  - Then, $P_{\theta} \to P$ as $\theta \downarrow 0$, we say it as P_{θ} **converges weakly**.
  - This weak convergence is known as P_{θ} **is tight**.

- If P_{θ} is tight, $f(x)$ **has a unique minimizer** on a given domain, when $f(x)$ is strongly convex around the global minimzer such that

$$\lim_{\theta\downarrow 0}\int_{\Omega} f(x)m^{\theta}(x)dx = \int_{\Omega} f(x)m^{0}(x)dx = f^{*} < f(x),\ \forall x \in \Omega$$

- Generally, Q is regareded as a normal distribution.
- P is the distribution derived by an optimization algorithm.
- Recently, Laplace method is said to be the **minima distribution** [2].

Stochastic Calculus

- To use Laplace's method, the analysis requires a **stochastic model for an algorithm**.
    - **Langevine** stochastic differential equation (SDE) model

$$dX_t = -\varepsilon \cdot \nabla f(X_t)dt + \sigma(t)\sqrt{G}dW_t$$

    - **Geometrical** stochastic differential equation (SDE) model

$$dX_t = -\varepsilon \cdot \nabla f(X_t)dt + \sigma \nabla f(X_t)dW_t$$

        - The coefficient of stochastic differential $\sqrt{G}$ in Langevine SDE appears as $\sqrt{\nabla f(x_t) \otimes \nabla f(x_t)} \sim \nabla f(x_t)$. (NTK or Natural Gradient?)

- Generally, **Langevine SDE model is appropriate** to analyze and implement.
    - Many algorithms for global optimization, such as **simulated annealing, quantum annealing, and genetic algorithm**, follows Langevine SDE.
        - **GAN**
        - **Baysian optimization**
        - **analysis of machine learning applied motion pictures**
    - By **Girsanov Theorem**, we can analyze the statistics and probability distribution of a learning algorithm.
- Geometrical SDE model easily analyzes the dynamics of the stochastic process.
    - Using Martingale analysis, we can analyze the dynamics of the random process itself.
    - Analysis of Momentum based Learning equation

$$X_t = X_0 \exp\left(-\varepsilon \nabla f(x) - \frac{\sigma^2}{2}\right) t + \sigma W_t$$

Girsanov Theorem

- Stochastic Calculus Version of **Radon-Nykodym Derivative**
- To obtain **the distribution of the continuous stochastic process** $\{X_t\}_{t=0}^{\infty}$ **represented with an SDE**.
  - There must be a **comparable stochastic process** such as **Wiener Process**.
  - It means that the probability of the process represented with an SDE process is a Radon-Nykodym derivative concerning the Wiener process with Gaussian distribution.

Example

- Suppose that the SDE of learning Equation is as follows:

$$dX_s = -\nabla f(X_s)ds + \sigma(s)\sqrt{G}dW_s,\ \ s \in \mathbf{R}(t, t+1).$$

- If the SDE of a standard Wiener Process is

$$d\bar{X}_s = \sigma(s)\sqrt{G}dW_s, \quad s \in \mathbf{R}(t, t+1),$$

- we get the distribution of the learning equation by Girsanov theorem, as follows:

$$\frac{dP_x}{dQ_x} = \exp\left\{-\int_t^{t+1} \frac{G^{-1}}{\sigma^2(s)}\nabla_x f(X_s), d\bar{X}_s - \frac{1}{2}\int_t^{t+1} \frac{G^{-1}}{\sigma^2(s)}\|\nabla_x f(X_s)\|^2 ds\right\}.$$

Frame work of Global Analysis based on Probability Space

- Fundamental Assumption
  - Lipschitz Continuous

$$\|\nabla f(w_s) - \nabla f(x^*)\| \leq L'\|w_s - x^*\|, \quad \forall s > 0.$$

- SDE of Learning equation

$$dX_s = -\nabla f(X_s)ds + \sigma(s)\sqrt{G}dW_s,\ s \in \mathbf{R}(t, t+1).$$

  - Set a standard Wiener process for the Girsanov theorem as follows:

$$d\bar{X}_s = \sigma(s)\sqrt{G}dW_s, \quad s \in \mathbf{R}(t, t+1).$$

- Girsanov Theorem

$$\frac{dP_x}{dQ_x} = \exp\left\{-\int_t^{t+1} \frac{G^{-1}}{\sigma^2(s)}\nabla_x f(X_s), d\bar{X}_s - \frac{1}{2}\int_t^{t+1} \frac{G^{-1}}{\sigma^2(s)}\|\nabla_x f(X_s)\|^2 ds\right\}.$$

- Analysis based on Calculus
    - Calculate supremum of the first and the second terms in the above Radon-Nykodym equation such that

$$\left\| \int_{t}^{t+1} \frac{G^{-1}}{\sigma(s)} \nabla_x f(X_s) d\bar{X}_s \right\| \leq \frac{C_1}{\sigma(s)}, \ \frac{1}{2} \left\| \int_{t}^{t+1} \frac{G^{-1}}{\sigma^2(s)} \|\nabla_x f(X_s)\|^2 ds \right\| \leq \frac{C_2}{2\sigma^2(s)}$$

, where C_1, C_2 is a constant by the analysis of an algorithm

- Apply the supremums of each term to the Girsanov equation such that

$$\frac{dP_w}{dQ_w} \geq \exp\left(-\frac{1}{\sigma(s)}\left(C_1 + \frac{C_2}{2\sigma(s)}\right)\right) \geq \exp\left(-\frac{C_3}{\sigma(s)}\right) \because C_3 > 2\sigma(0)C_2 + C$$

- Consequently, for any $\varepsilon > 0$ and $x_t,\ x^* \in \mathbf{R}^n$, the infimum of $P_x(|X_{t+1} - x^*| < \varepsilon)$ is

$$P_x(|X_{t+1} - x^*| < \varepsilon) \geq \exp\left(-\frac{C_3}{\sigma(s)}\right) Q_x(|X_{t+1} - x^*| < \varepsilon).$$

where Q_x is a gaussian distribution by the standard Wiener process assumption.

- Proof of Convergence with Laplace's Method
  - Let the infimum of the transition probability from t to $t+1$ such that

$$\begin{aligned}
&\inf_{x,y\in\mathbf{R}^n} p(t,x,t+1,y)\Big|_{x=x_t,\ y=x^*} \\
&= \inf_{x,y\in\mathbf{R}^n} \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} P_x(|X_{t+1}-x^*|<\varepsilon) \\
&\geq \inf_{x,y\in\mathbf{R}^n} \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\cdot C_4\cdot \exp\left(-\frac{C_3}{\sigma(s)}\right) Q_x(|X_{t+1}-x^*|<\varepsilon).
\end{aligned}$$

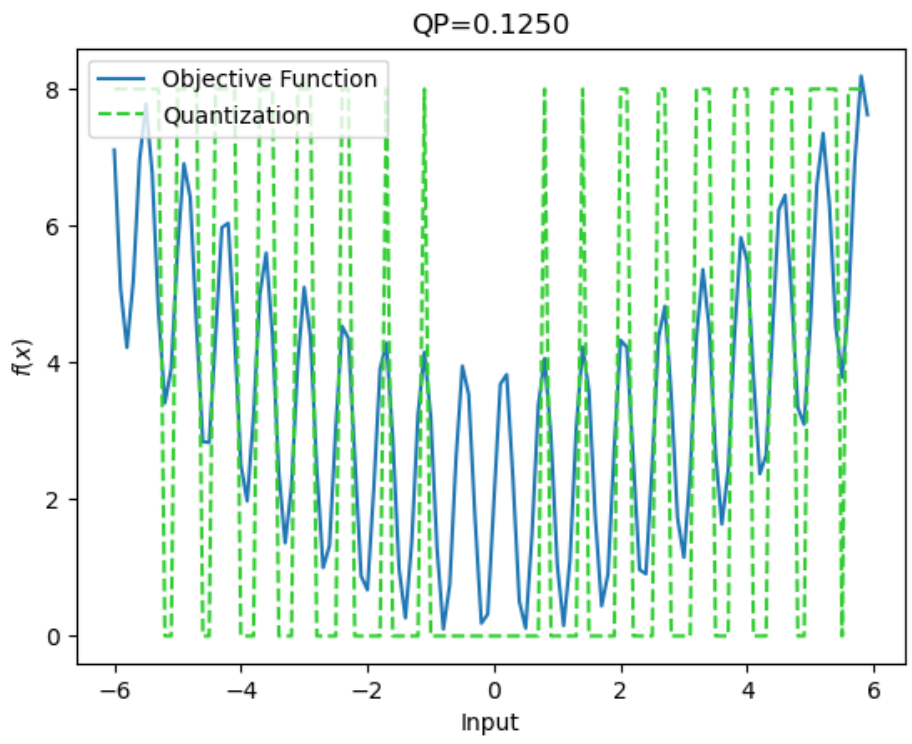
  - Evaluate the infimum of Q_x deriven by an algorithm.
  - With a limitation lemma of the difference to the transition probability such that

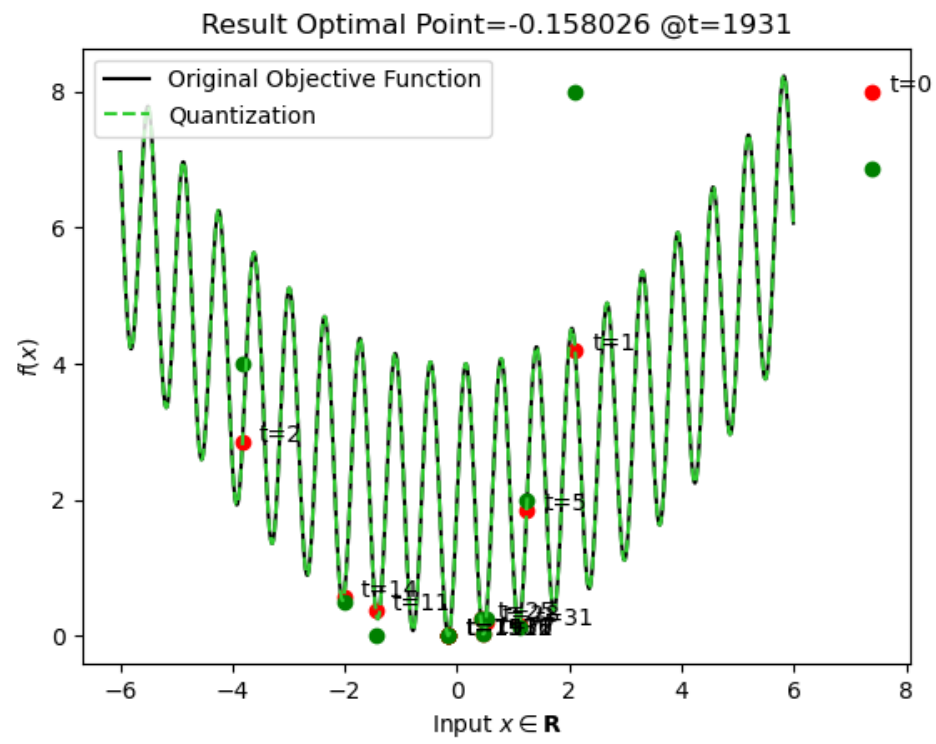
$$\begin{aligned}
&\lim_{t\to\infty} \sup_{w\in[0,1]^n} |p(s,v,t,x^*)-p(s,w,t,x^*)| \\
&= 2\cdot\|x^*\|_\infty \prod_{k=0}^{\infty}(1-\inf_{x,y\in\mathbf{R}^n} p(s+k,x_t,s+k+1,x^*)), \\
&\forall x\in D_0\subset\mathbf{R}^n, s\geq 0
\end{aligned}$$

, we can prove the global convergence of an algorithm.

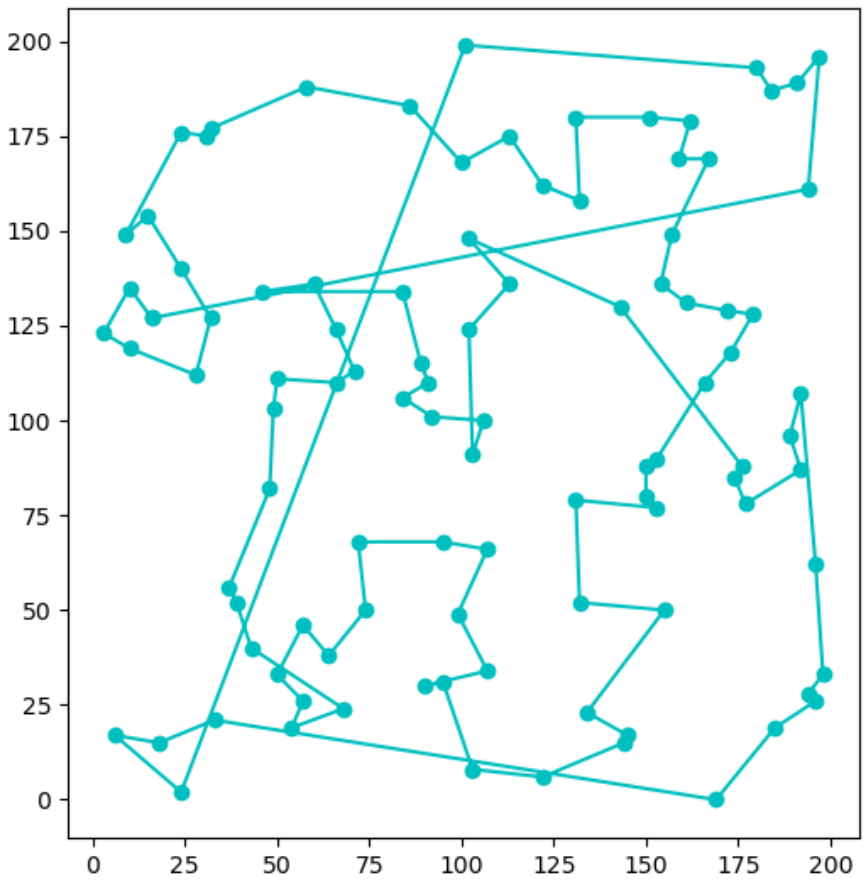
Examples : Global Optimization

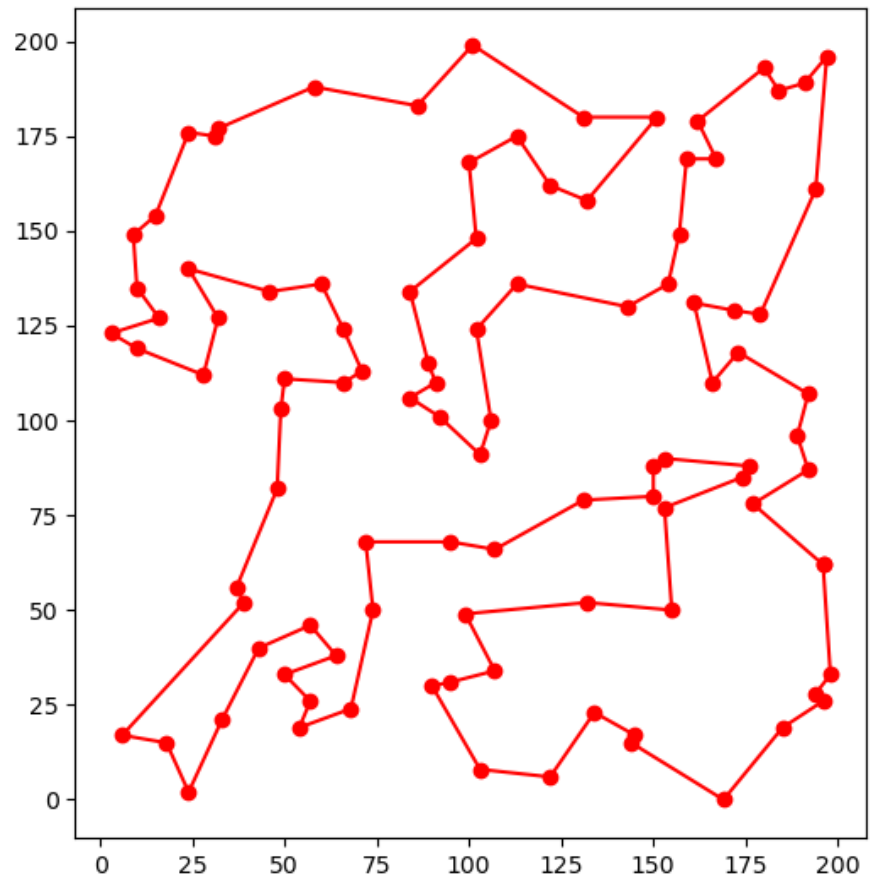
- Parabolic Washboard Potential Function
    - The test function for comparison between quantum annealing and simulated annealing





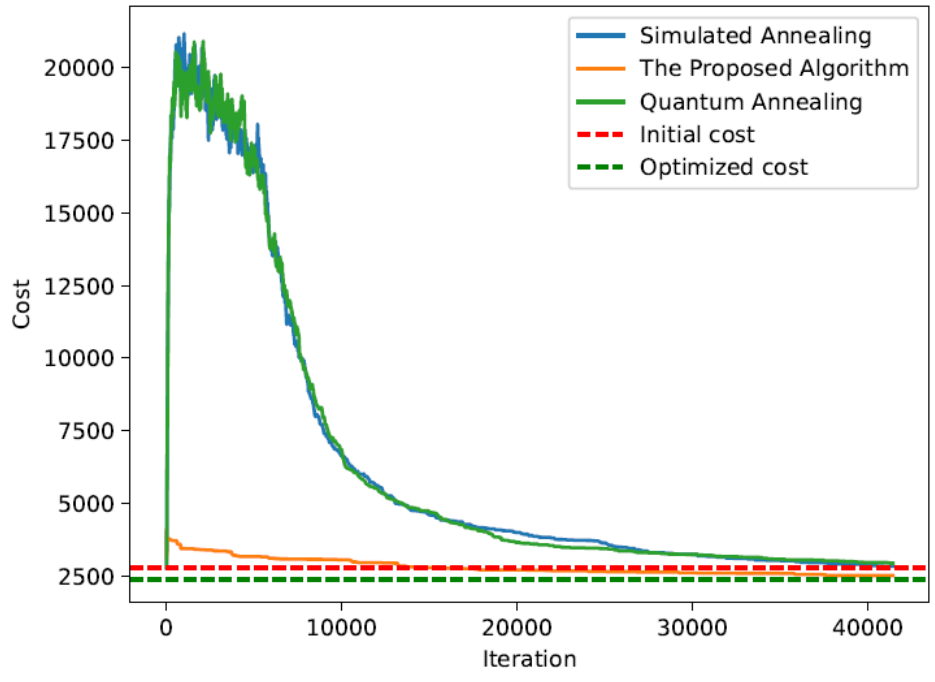
- Travelling Salesman Problem
  - Representative Np-Hard Problem
  - The algorithm **without the proof of global optimization cannot show better performance than the Nearest Neighborhood algorithm**.





Conclusion

cities	NN(Initial)	SA	QA	QZ	Improve ratio
100	2159.27	1727.44	1729.69	1706.53	20.96
125	2297.86	2027.52	2028.2	1923.65	16.28
150	2497.65	2255.15	2252.82	2032.21	18.63
175	2380.52	2380.52	2380.29	2147.17	9.80
200	2769.73	2769.34	2769.42	2366.72	14.55



- The proof of global optimization is **the proof of optimization after Hill-Climbing**.
  - The conventional analysis of Hilbert space provides the **divergence of the algorithm involving the Hill-Climbing**.
  - Additionally, the proof of **asymptotic convergence requires diminishment of a Hill-Climbing effect**.

- Development of an algorithm for global optimization
    - Conventional research of global optimization is based on **Monte-Carlo sampling**
        - Or **additional noise such as GAN and Bayesian Optimization**
    - Such as quantization or quantum computing, using uncertainty in the domain may be effective for global optimization. - Optimization algorithm based on number theory.
- Development of global optimization on the differential manifold.
    - There is a lot of new research after failure on stochastic filtering on the differential manifold.
    - Requirement of Wiener Proces on tangent space for learning equation.
    - Researching the global optimization based on a combination of gradient replacing a Hessian
        - NTK and Natural Gradient are so effective? Sufficient?

References

[1] Chii-Ruey Hwang. "Laplace's Method Revisited: Weak Convergence of Probability Measures." Ann. Probab. 8 (6) 1177 - 1182, December, 1980. https://doi.org/10.1214/aop/1176994579

[2] Xiaopeng Luo, "Minima distribution for global optimization", 2018, arXiv:1812.03457

[3] Seok, J. and Kim, J.-S., Nonlinear optimization algorithm using monotonically increasing quantization resolution, *ETRI Journal* (2022), 1– 12. https://doi.org/10.4218/etrij.2021-0320

Thank you