Model-driven Learning to be Global Optimizer (Proof)

Abstract

We summarize the proof of the paper, the demonstrations are the same in the paper.

1 Proof of Gd-Net

In the following we show that Gd-Net is guaranteed to local optimality. We first prove that HGD is convergent. Theorem 1 shows the result.

Theorem 1 *Assume $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and differentiable and the sublevel set*

$$L(x) = \Big\{x \in \mathbb{R}^n | f(x) \leq f(x_0)\Big\} \text{ for any } x_0 \in \mathbb{R}^n$$

is bounded. The sequence $\{x_k, k = 1, 2, \cdots\}$ obtained by HGD with exact search converges to a local minimum $x^\star$.

Since d-Net is the unfolding of HGD, from Theorem 1, it is sure that the sequence obtained by d-Net is non-increasing for any initial x_0. Gd-net is a sequence of d-Net. Applying Gd-Net on a bound function $f(x)$ from any initial point x_0 will result in a sequence of function values which is non-increasing. This ensures that the sequence will converge, which indicates that Gd-Net is convergent under the assumption of Theorem 1.

Proof 1 *In HGD, $d_k = -R_{k-1}\widetilde{H}_{k-1}g_k$ where*

$$R_{k-1} = \left[I - \frac{s_{k-1}\widehat{y}_{k-1}^\intercal}{s_{k-1}^\intercal \widetilde{y}_{k-1}}\right]$$

and $\widetilde{y}_{k-1} = w_k^3 g_k - w_k^4 g_{k-1}, \widehat{y}_{k-1} = w_k^1 g_k - w_k^2 g_{k-1}, \widetilde{H}_{k-1} = \beta_k H_{k-1} + (1-\beta_k)I$. With exact linear search, we have $g_k^\intercal s_{k-1} = 0$, therefore

$$\begin{aligned} -g_k^\intercal d_k &= g_k^\intercal(\beta_k H_{k-1} + (1-\beta_k)I)g_k \\ &- g_k^\intercal \frac{s_{k-1}\widehat{y}_{k-1}^\intercal}{s_{k-1}^\intercal \widetilde{y}_{k-1}} \widetilde{H}_{k-1} g_k \\ &= g_k^\intercal(\beta_k H_{k-1} + (1-\beta_k)I)g_k \end{aligned}$$

It is clear that $-g_k^\intercal d_k > 0$ if $H_{k-1} \succ 0$, otherwise a $\beta_k > 0$ can be chosen to make $(\beta_k H_{k-1} + (1-\beta_k)I)$ diagonally dominant, which means $g_k^\intercal(\beta_k H_{k-1} + (1-\beta_k)I)g_k > 0$. Therefore, we can always make sure $g_k^\intercal d_k < 0$, i.e. d_k is a descent direction, and

$$f(x_1) \geq f(x_2) \geq \cdots \geq f(x_k)$$

Since $f(x)$ is bounded, there exists $f(x^*)$ such that $\lim_{k\to\infty} f(x_k) = f(x^\star)$. □

2 Proof of reward

Theorem 2 *If $x' = x_0 + T \cdot d$ is a point outside the boundary of x_0's attraction basin, i.e. $x' \notin \partial\mathcal{B}(x_0, \delta)$ and there is no other local minima within $\mathcal{B}_0 = \{x \in \mathbb{R}^n | \|x - x_0\|_2 \leq \|x' - x_0\|_2\}$. Then*

$$g(t) \doteq f(x_0 + t \cdot d), t \in [0, T]$$

has a maximum ξ. And $g(t)$ is monotonically increasing in $[0, \xi]$, and monotonically decreasing in $(\xi, T]$.

Proof 2 *According to assumption (2)(3), $g(t)$ is convex in $[-\delta, \delta]$. As $g'(0) = f'(x)\big|_{x=x_0} = 0, g''(0) > 0$, then $g'(t) > 0$ in $(0, \delta]$. Therefore, $g(t)$ is monotonically increasing in $[0, \delta]$, and monotonically decreasing in $[T-\delta, T]$. Since $f(x) \in C^2(\mathbb{R}^n)$, then $g(t) \in C^2[0, T]$. This implies that $g'(t)$ is continuous in $[0, T]$.*

Since $g'(\delta) > 0$ and $g'(T-\delta) < 0$, then there exists a ξ such that $g'(\xi) = 0$. Further, ξ is unique since it is assumed that there is no other local minima between x_0 and x_1. Then we have $g'(t) > 0$ in $[0, \xi]$, and $g'(t) < 0$ in $[\xi, T]$. □

3 Proof of step size

First, we shall prove the existence of

$$a^* = \underset{a \in \mathbb{R}_{++}}{\operatorname{argmin}} F(a). \tag{1}$$

along d in case there exists an x' and $f(x') < f(x_0)$.

Lemma 1 *Suppose $F(a)$ to be the function defined in Alg.2. For fixed a and N, we have:*

$$\lim_{\alpha \to 0} F'(a) = \frac{1}{a}\int_{t_1}^{t_N} g'(t)dt \tag{2}$$

Proof 3 *Since $\{x_1, \cdots, x_N\}$ is obtained by applying gradient descent on $\widetilde{H}(x)$ along direction d, they are all on the line $x_0 + td$ with $t > 0$. Therefore, each $x_i, 1 \leq i \leq N$ can be written as*

$$x_i = x_0 + t_i d$$

where $t_1 = \delta > 0$ and $t_1 < t_2 \cdots < t_N$. Further, we have

$$x_i = x_{i-1} - \alpha \nabla \widetilde{H}(x) = x_{i-1} + \alpha \cdot 2a(x_{i-1} - x_0)$$

where α is the step size. It can be re-written as

$$x_i = x_{i-1} + a\tilde{\alpha}\cdot d \quad \Leftrightarrow \quad x_i - x_{i-1} = a\cdot 2\alpha t_{i-1}\cdot d$$
$$\Leftrightarrow \quad (t_i - t_{i-1})\cdot d = a\cdot 2\alpha t_{i-1}\cdot d$$

or equivalently,

$$t_i - t_{i-1} = a\cdot 2\alpha t_{i-1} \Longrightarrow t_i = (1+2\alpha a)^{i-1}t_1 \qquad (3)$$

We thus have:

$$\begin{aligned} F'(a) &= \sum_{i=1}^{N}\nabla_a f(x_i) = \sum_{i=2}^{N}\nabla_a f(x_{i-1} + a\cdot 2\alpha t_{i-1}d) \\ &= \sum_{i=2}^{N}\nabla f(x_{i-1} + a\cdot 2\alpha t_{i-1}d)^\intercal \cdot 2\alpha t_{i-1}d \\ &= \sum_{i=2}^{N}\nabla f(x_{i-1} + a\cdot 2\alpha t_{i-1}d)^\intercal \cdot \frac{1}{a}(t_i - t_{i-1})d \\ &= \frac{1}{a}\sum_{i=2}^{N} g'(t_i)(t_i - t_{i-1}) \end{aligned}$$

Since $g'(t)$ is continuous in $[t_1, t_N]$, it is Riemann integrable. Thus, for a fixed N, we have

$$\lim_{\alpha\to 0}\sum_{i=2}^{N} g'(t_i)(t_i - t_{i-1}) = \int_{t_1}^{t_N} g'(t)dt$$

□

In the sequel, we define

$$G(a) = t_N - t_1 = ((1+2\alpha a)^{N-1} - 1)t_1$$

where $a \in [0, \infty)$. Here $G(a)$ is just the distance between x_N and x_1 along d. Obviously, $G(a)$ is a polynomial function of a, and its is monotonically increasing.

Theorem 3 *Suppose that $x' = x_0 + Td$ is a point such that $f(x_0) \geq f(x')$, and there are no other local minima within $\mathcal{B}_0$. If α is sufficiently small, then there exists an a^* such that $F'(a^*) = 0$.*

Proof 4 *Since $f(x') \leq f(x_0)$, and $g(t) = f(x_0 + td)$ is smooth, there exists a ξ such that $\xi = \text{argmax}_{t\in[t_1,T]}\, g(t)$.*

Let's consider two cases. First, let $D_1 = \xi - t_1$ i.e. all $x_i \leq \xi$, then there is an a_1 s.t. $G(a_1) = D_1$ according to the intermediate value theorem. The derivative of $F(a_1)$ can be computed as follows:

$$F'(a_1) = \sum_{i=1}^{N}\nabla_a f(x_i)\Big|_{a_1} = \sum_{i=2}^{N}\nabla_a f(x_{i-1} + a\tilde{\alpha}d)\Big|_{a_1}$$
$$= \sum_{i=2}^{N}\widetilde{\alpha}\nabla f(x_{i-1} + a_1\tilde{\alpha}d)^\intercal \cdot d = \sum_{i=2}^{N}\widetilde{\alpha}\nabla f(x_i)^\intercal \cdot d > 0$$

where the last equality holds because $\nabla f(x_i)^\intercal \cdot d > 0$ for all $t_i < \xi$.

Similarly, if let $D_2 = T - t_1$, then there is a_2 s.t. $G(a_2) = D_2$, as α is sufficiently small, we have: $\forall\varepsilon > 0,\ \exists\,\alpha$ s.t.

$$\left|F'(a_2) - \frac{1}{a_2}\int_{t_1}^{t_N} g'(t)dt\right| < \varepsilon$$

Note that

$$\frac{1}{a_2}\int_{t_1}^{t_N} g'(t)dt = \frac{1}{a_2}(g(T) - g(t_1)) \quad =$$
$$\frac{1}{a_2}[f(x') - f(x_0 + \delta d)] = \frac{1}{a_2}(f(x') - f(x_1)) \quad < \quad 0$$

Let $\epsilon_0 = \frac{1}{a_2}\int_{t_1}^{t_N} g'(t)dt$ and $\epsilon = \epsilon_0/2$, then $\exists\ \alpha_0$ s.t. $|F'(a_2) - \epsilon_0| < -\epsilon_0/2 \ \Rightarrow\ F'(a_2) < \epsilon_0/2 < 0$.

In summary, we have $F'(a_1) > 0$ and $F'(a_2) < 0$, according to the intermediate value theorem, there exists an a^* such that $F'(a^*) = 0$. □

According to this theorem, we have the following corollary.

Corollary 1 *Suppose that x_N is the solution obtained by optimizing $\widetilde{H}(x) = -a^*\|x - x_0\|_2^2$ along d starting from $x_0 + \delta d$ at the N-th iteration, then x_N will be in an attraction basin different from x_0, if the basin ever exists.*

In Alg.2, for a given a, a series of points $x_2, \cdots, x_N$ is obtained by minimizing $\widetilde{H}(x)$. If at some k,

$$g(t_k) \approx \max f(x_0 + td),$$

then $g'(t_k) \approx 0$. According to Theorem.2, we know that $g'(t_i) > 0$ for $1 \leq i \leq k-1$. Therefore $\sum_{i=1}^{k} g'(t_i) > 0$. If the sum of the rest gradients $\sum_{j=k+1}^{N} g'(t_j)$ cannot compensate $\sum_{i=1}^{k} g'(t_i)$, i.e.

$$\sum_{j=k+1}^{N} g'(t_j) + \sum_{i=1}^{k} g'(t_i) \neq 0$$

then a is not a stable point, which means that a larger a is required.

4 Proof of the sampling method

We prove that the developed sampling is more efficient than that of the random sampling (i.e. directions are sampled randomly from the search space) in terms of finding a promising direction. Theorem 4 summarizes the results. In the theorem, denote P_r be the probability of finding a promising direction by using the random sampling, P_c be the probability by Alg.3. We first define 'opposite cone B^*'.

Definition 1 *Given a set of directions $\{d_1, \cdots, d_N\}, N \ll n$, the opposite cone B^* is defined as*

$$B^* = \{\alpha_1 d_1 + \cdots + \alpha_N d_N | - \alpha_i \in \mathbb{R}_{++}, d_i \in \mathbb{R}^n\}$$

Theorem 4 *Given an initial sample of directions and rewards $\{(d_1, r_1)\cdots, (d_N, r_N)\}$, the direction sampled by using Alg.3 is of higher probability to sampling directions with positive rewards (or promising) than that of the random sampling, i.e. $P_c > P_r$.*

Proof 5 *Without loss of generality, suppose that we are at a local minimum x_0, and try to sample a direction that points to another local minimum x'. Denote the direction as $d^* = x' - x_0$. Let $\Omega = \{x : \|x - x_0\|_2 \leq M\}$ be the confined search space.*

First, when $N < n$, then N linear independent vectors can result in a 'cone'. These vectors then divide the search space into 2^N cones $B_1, B_2, \cdots, B_{2^N}$. Suppose that x' has an attraction basin, denoted as $R_{x'}$, and the radius of the inscribed sphere of $R_{x'}$, is r'. Then the boundary of $B(x', r')$ and x_0 can form a cone C^*. Suppose the lines $x_0 + td_i, i = 1, \cdots, N$ has no interaction with C^* (otherwise we would have found a direction that will lead to the attraction basin of x'), i.e.

$$C^* \bigcap \{x | x = x_0 + td_i, x \in \Omega\} = \emptyset, \forall i \in \{1, ..., N\}, t > 0$$

For each d_i, take x^i such that $x^i \in \partial\Omega \bigcap d_i$. Then the boundary of $B(x^i, r_1)$ and x_0 form a cone C^i, let $\widetilde{C} = \bigcup_{i=1}^{N} C^i$, we have

$$\widetilde{C} \bigcap \{x | x = x_1 + td^*, x \in \Omega\} = \emptyset, t > 0$$

which implies that within the cones $C^i, 1 \leq i \leq N$, there are no promising directions. Thus, we should avoid looking for directions in the union of C^is.

Notice that $B^* \bigcap C^i = \emptyset$ for $i = 1, \cdots, N$. Denote

$$\widetilde{\Omega} = \left\{ x | x \in \Omega, x \notin \widetilde{C} \right\}$$

Then P_c, the probability of finding a promising direction in Ω by reinforcement sampling, can be computed as follows:

$$\begin{aligned} P_c &= P\{x_1 \in B^*\} P\{d \text{ is promising } | x_1 \in B^*\} \\ &= \frac{V_{B^*}}{V_{\widetilde{\Omega}}} \cdot P_r \cdot \frac{V_\Omega}{V_{B^*}} = \frac{V_\Omega}{V_{\widetilde{\Omega}}} \cdot P_r > P_r \end{aligned}$$

where $V_{B^*}, V_\Omega, V_{\widetilde{\Omega}}$ is the volume of B^*, Ω and $\widetilde{\Omega}$, respectively. The last inequality holds because $\widetilde{\Omega} \subset \Omega$. □

Fig. 1 illustrates the proof in 2-D case, in which it is clear that sampling in B^* is the wisest choice for a direction with potentially positive rewards.

Note: Theorem 4 proves that statistically we can use sample less times to find another local minima than random sampling. In case $f(x)$ has 3 or more local minima, the sampling procedure can be done as follows. Assuming we have sampled P_0 directions, $\{d_i\}_{i=1}^{P_0}$, from which suppose at least one local minima x_{last} cannot be reached. It is therefore not wise to sample within the cones induced by these local minima. Instead, the negative rewards associated with these directions should be used as the linear combination for sampling x_{last}, i.e. line 7 of Alg.3 can be used. Theorem 4 again can be used to guarantees that this combination is more efficient to create promising directions for x_{last} than random sampling.

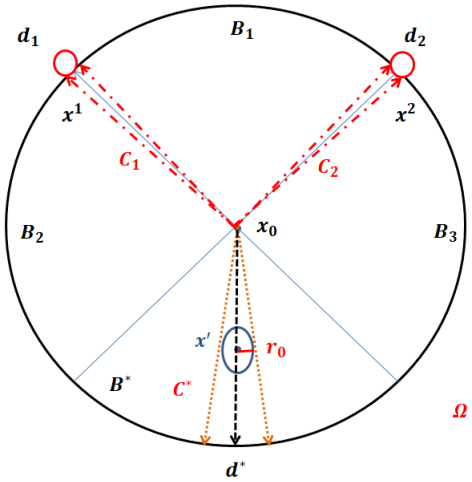

Figure 1: Illustration of the proof of Theorem 4 in 2-D case. In the figure, d^* is to be found in Ω. r_0 is the radius of the inscribed sphere of the attraction basin of x'. $x_0, B(x', r_0)$ form a cone C^*. For each d_i, take $x^i \in \partial\Omega \bigcap d_i$, the boundary of $B(x^i, r_0)$ and x_0 forms a cone C^i. Then $C^i \bigcap \{x | x = x_1 + td^*, x \in \Omega\} = \emptyset, t > 0, \forall i$. It is clear that sampling a direction in B^* is the wisest choice.