National Seminar on
Control and Optimization
Nov 13-14, 2019
3th
سومین سمینار ملی
کنترل و بهینه‌سازی
۲۳-۲۲ آبان ۱۳۹۸

ارائه‌ی الگوریتمی جدید برای یادگیری عمیق مبتنی بر مسئله کنترل بهینه گسسته-زمان

نویسنده دکتر محمود محمودی*، نویسنده سارا احمدی[۲]

گروه ریاضی کاربردی، دانشکده علوم پایه، دانشگاه قم

mmahmoodi_ir@yahoo.com

[۲]گروه ریاضی کاربردی، دانشکده علوم پایه، دانشگاه قم

sara.ahmadi1389@yahoo.com

چکیده: در این مقاله یادگیری عمیق به‌عنوان یک مسئله کنترل بهینه گسسته-زمان فرموله می‌شود. این امر ما را قادر می‌سازد که شرایط لازم برای بهینگی را به شکلی خاص مشخص کرده و الگوریتم‌های یادگیری‌ای که متکی به گرادیان نسبت به پارامترهای آموزش‌پذیر هستند را توسعه دهیم. در حالت خاص روش تقریب‌های متوالی گسسته-زمان (MSA) را معرفی می‌کنیم که بر اساس اصل ماکسیمم پونتریاگین (PMP) برای شبکه‌های عصبی آموزش‌پذیر پایه‌ریزی شده‌اند. علاوه بر این یک تقریب خطای بسیار دقیق برای MSA گسسته بدست می‌آید که می‌تواند برای بدست آوردن الگوریتم‌های مفید استفاده شود. محاسبه‌ی این تقریب خطا را می‌توان به‌عنوان نتیجه اصلی این مقاله از نظر تئوری تلقی کرد.

۱. پیش‌گفتار

مسئله آموزش عمیق شبکه‌های عصبی پیشخور[†] اغلب به‌عنوان یک مسئله برنامه‌ریزی غیرخطی نامقید مورد مطالعه قرار می‌گیرد. در این مسائل که اغلب مینیمم سازی تابع هدف، روی پارامترهای آموزش‌پذیر انجام میگیرد، شرایط لازم برای بهینگی معمولا توسط برابر صفر قرار دادن گرادیان تابع هدف نسبت به متغیرهای آموزش‌پذیر، صورت می‌گیرد. این امر تا حد زیادی پایه‌ای برای الگوریتم‌های بهینه‌سازی مبتنی بر گرادیان کاهشی در یادگیری عمیق است. هنگامی که محدودیت‌های اضافی روی پارامترهای آموزش‌پذیر وجود دارد، می‌توان از نسخه‌های پیش‌بینی شده الگوریتم فوق استفاده کرد. به‌طور گسترده‌تر شرایط لازم برای بهینگی می‌تواند در قالب شرایط KKT حاصل شود. چنین رویکردهایی کاملا کلی نیست و به‌طور معمول به ساختار اهداف مورد نظر در یادگیری عمیق متکی نیست. با این حال در یادگیری عمیق تابع هدف اغلب از ساختار خاصی برخوردار است. این ساختار از تغذیه دسته‌ای از ورودی‌ها به صورت بازگشتی از طریق دنباله‌ای از تبدیلات قابل آموزش حاصل می‌شود که می‌توان آن‌ها را طوری تنظیم کرد تا خروجی‌های نهایی نزدیک به برخی از مجموعه اهداف ثابت باشند. این پروسه شبیه به یک مسئله کنترل بهینه است که از مطالعه حساب تغییرات ایجاد می‌شود. در این مقاله از دیدگاه کنترل بهینه از یاد گیری عمیق رونمایی می‌شود.

واژگان کلیدی: یادگیری عمیق، مسئله کنترل بهینه، روش تقریب‌های متوالی، اصل ماکسیمم پونتریاگین

* سخنران

† Feed-Forward Neural Networks

۰۲ دیدگاه کنترل بهینه

در این بخش مسئله آموزش یک شبکه عصبی عمیق را به‌عنوان یک مسئله کنترل بهینه رسمیت می‌دهیم. فرض کنید $T \in Z_+$ تعداد لایه‌ها و $\{x_{s,0} \in R^{d_0} : s = 0, \dots, S\}$ مجموعه‌ی ورودی‌های ثابت شبکه را نشان دهد که در آن $S \in Z_+$ اندازه‌ی نمونه است. سیستم دینامیکی زیر را در نظر بگیرید:

$$x_{s,t+1} = f_t(x_{s,t}, \theta_t), \qquad t = 0,1, \dots, T-1 \qquad (1)$$

که در آن $f_t: R^{d_t} \times \Theta_t \longrightarrow R^{d_{t+1}}$ به ازای هر t، یک تبدیل روی متغیر حالت است. فرض کنید که هر مجموعه پارامتر آموزش‌پذیر Θ_t یک زیر مجموعه از فضای اقلیدسی است. هدف از آموزش یک شبکه عصبی تنظیم وزن‌های $\{\theta_t : t = 0, \dots, T-1\}$ به‌گونه‌ای است که تابع هزینه‌ای را حداقل کند که در آن اختلاف بین خروجی نهایی شبکه یعنی $x_{s,T}$ و مقادیر واقعی هدف یعنی y_s لحاظ شده است. برای این منظور خانواده‌ای از توابع حقیقی مقدار $\Phi_s: R^{d_T} \to R$ که روی $x_{s,T}$ عمل می‌کنند تعریف می‌کنیم. علاوه بر این ممکن است به‌ازای هر لایه تابع نظم‌دهی[‡] $L_t: R^{d_t} \times \Theta_t \longrightarrow R$ را در نظر بگیریم که باید همزمان با تابع هزینه حداقل شود. بنابراین هدف ما در شبکه عصبی حل مسئله زیر است:

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Phi}} \mathrm{J}(\boldsymbol{\theta}) := \frac{1}{S}\sum_{s=1}^{S} \Phi_s(x_{s,T}) + \frac{1}{S}\sum_{s=1}^{S}\sum_{t=0}^{T-1} L_t(x_{s,t}, \theta_t) \qquad (2)$$

$$s.t. \qquad x_{s,t+1} = f_t(x_{s,t}, \theta_t), \qquad t = 0,1, \dots, T-1 \ , s \in [S]$$

که خلاصه کرده‌ایم: $\Phi := \{\Phi_0, \dots, \Phi_{T-1}\}$ و $\{1,2,\dots,S\}=:[S]$.

۲-۱ اصل ماکسیمم پونتریاگین

اصل ماکسیمم پونتریاگین[§] (PMP) معمولا از شرایط لازم برای بهینگی در قالب حداکثر رساندن یک تابع خاص همیلتون تشکیل شده است. از ویژگی‌های بارز این اصل این است که فرض مشتق‌پذیری (ویا حتی پیوستگی) تابع f_t نسبت به θ دیگر لازم نیست. بنابراین شرط بهینگی و الگوریتم‌های مبتنی بر آن نیازی به به‌روزرسانی روش‌های گرادیان کاهشی ندارد. این یک ویژگی خاص برای رده‌ای از برنامه‌هاست. فرض کنید $\boldsymbol{\theta}^* = \{\theta_0, \dots, \theta_{T-1}\} \in \boldsymbol{\Theta}$ جواب مسئله (۲) باشد ، اکنون به‌طور غیر رسمی PMP که $\boldsymbol{\theta}^*$ را مشخص می‌کند را تشریح می‌کنیم. ابتدا برای هر t تابع همیلتونی $R_t: R^{d_{t+1}} \times \Theta_t \longrightarrow R$ را بوسیله رابطه زیر تعریف می‌کنیم

$$H_t(x, p, \theta) := p.f_t(x, \theta) - \frac{1}{S}L_t(x, \theta) \qquad (3)$$

اکنون شرایط لازم زیر را می‌توان ثابت کرد.

قضیه ۱.۲ (PMP گسسته) فرض کنید f_t و Φ_s، s=1,…,S به اندازه کافی در x هموار باشند. همچنین فرض کنید که برای هر t و $x \in R^{d_t}$ ، مجموعه‌های $\{f_t(x, \theta) : \theta \in \Theta\}$ و $\{L_t(x, \theta) : \theta \in \Theta\}$ محدب باشند. آن‌گاه متغیرهای شبه‌حالت $\boldsymbol{p}_s^* := \{p_{s,t}^* : t = 0, \dots, T\}$ وجود دارند به‌طوری که شرایط زیر برای t=0,…,T−1 و $s \in [S]$ برقرار هستند:

[‡] regularization terms

[§] The Pontryagin's Maximum Principle

$$x^*_{s,t+1} = \nabla_p H_t\left(x^*_{s,t}, p^*_{s,t+1}, \theta^*_t\right), \quad x^*_{s,0} = x_{s,0} \qquad (4)$$

$$p^*_{s,t} = \nabla_x H_t\left(x^*_{s,t}, p^*_{s,t+1}, \theta^*_t\right), \quad p^*_{s,T} = -\frac{1}{S}\nabla\Phi_s(x^*_{s,T}) \qquad (5)$$

$$\sum_{s=1}^{S} H_t\left(x^*_{s,t}, p^*_{s,t+1}, \theta^*_t\right) \geq \sum_{s=1}^{S} H_t\left(x^*_{s,t}, p^*_{s,t+1}, \theta\right), \quad \forall\theta \in \Theta_t \qquad (6)$$

برهان. به [1] مراجعه شود.

بگذارید PMP رابا جزئیات شرح دهیم. معادله حالت (۴) انتشار به‌جلو[**] معادله (۱) تحت پارامتر بهینه $\boldsymbol{\theta}^*$ است. معادله (۵) سیر تکاملی متغیر شبه حالت $\boldsymbol{p}^*_s$ را تعریف می‌کند. شاید مناسب‌تر باشد که از متغیرهای $p_{s,t}$ در معادله (۵) به‌عنوان سیر تکاملی بردار نرمال یک ابرصفحه یاد کنیم که مجموعه‌ای از حالت‌های قابل دستیابی و مجموعه حالت‌هایی را که مقدار تابع هدف آن‌ها کوچک‌تر از مقدار بهینه است را از هم جدا می‌کند.(برای تحلیل بیشتر پیوست A از [1] را ببینید). شرط حداکثر رساندن تابع هامیلتونین (معادله ۶) قسمت اصلی PMP می‌باشد. معادله (۶) بیان می‌کند که جواب بهینه $\boldsymbol{\theta}^*$ باید به‌طور سراسری تابع هامیلتونین را برای هر لایه t=0,…,T−1 ماکسیمم کند.

۳. روش تقریب‌های متوالی

PMP (معادلات ۶-۴) یک مجموعه از شرایط لازم را به ما می‌دهد که یک جواب بهینه برای (۲) باید آن‌ها را صدق دهد. با این حال به ما نمی‌گوید که چنین جوابی را چگونه باید پیدا کرد. هدف این بخش بحث در مورد الگوریتم‌هایی برای حل (۲) بر اساس اصل ماکسیمم است. در بررسی دقیق از معادلات ۶-۴ می‌توان در یافت که هر یک از آن‌ها یک منیفلد در فضای جواب شامل همه $\boldsymbol{\theta}$ های ممکن، $\{\boldsymbol{x}_s,\ s \in [S]\}$ و $\{\boldsymbol{p}_s,\ s \in [S]\}$ را نمایش می‌دهند و اشتراک این سه منیفلد در صورت وجود، می‌تواند شامل یک جواب بهینه باشد. در نتیجه یک روش تکراری که یک جواب حدسی را به توی هریک از این منیفلدها تصویر می‌کند می‌تواند مفید باشد. این روش، روش تقریب‌های متوالی[††] (MSA) است که در ابتدا برای حل مسائل کنترل بهینه پیوسته-زمان معرفی شد. اجازه دهید اکنون نسخه گسسته-زمان را معرفی کنیم.

- با یک جواب حدسی $\boldsymbol{\theta}^0 = \{\theta^0_t, t = 0, \dots, T-1\}$ شروع کنید.
- برای هر نمونه s با استفاده از $x^{\boldsymbol{\theta}^0}_{s,t+1} = f_t(x^{\boldsymbol{\theta}^0}_{s,t}, \theta^0_t)$ برای t=0,…T-1 تعریف کنید $x^{\boldsymbol{\theta}^0}_s :=$ $\left\{x^{\boldsymbol{\theta}^0}_{s,t} : t = 0, \dots, T\right\}$. به طور شهودی این یک نمایش بر روی منیفلد تعریف شده توسط معادله (۴) است.
- سپس تصویرسازی را روی منیفلد تعریف شده توسط معادله (۵) انجام می‌دهیم یعنی تعریف می‌کنیم $p^{\boldsymbol{\theta}^0}_s :=$ $\left\{p^{\boldsymbol{\theta}^0}_{s,t}, \theta^0_t\right\}$ بوسیله متغیرهای $p^{\boldsymbol{\theta}^0}_{s,t} = \nabla_x H_t\left(x^{\boldsymbol{\theta}^0}_{s,t}, p^{\boldsymbol{\theta}^0}_{s,t+1}, \theta^0_t\right)$ برای t=0,…,T−1 و $p^{\boldsymbol{\theta}^0}_{s,T} =$ $-\frac{1}{S}\nabla\Phi_s\left(x^{\boldsymbol{\theta}^0}_{s,T}\right)$.

[**] forward propagation
[††] The Method of Successive Approximations

Algorithm 1 Basic MSA

Initialize: $\boldsymbol{\theta}^0 = \{\theta_t^0 \in \Theta_t : t = 0 \dots, T-1\}$;
for $k = 0$ **to** #Iterations **do**
$x_{s,t+1}^{\boldsymbol{\theta}^k} = f_t(x_{s,t}^{\boldsymbol{\theta}^k}, \theta_t^k), x_{s,0}^{\boldsymbol{\theta}^k} = x_{s,0}, \forall s, t;$
$p_{s,t}^{\boldsymbol{\theta}^k} = \nabla_x H_t(x_{s,t}^{\boldsymbol{\theta}^k}, p_{s,t+1}^{\boldsymbol{\theta}^k}, \theta_t^k), p_{s,T}^{\boldsymbol{\theta}^k} = -\frac{1}{S}\nabla\Phi_s(x_{s,T}),$
$\forall s, t;$
$\theta_t^{k+1} = \arg\max_{\theta\in\Theta_t} \sum_{s=1}^{S} H_t(x_{s,t}^{\boldsymbol{\theta}^k}, p_{s,t+1}^{\boldsymbol{\theta}^k}, \theta)$ for $t =$
$0, \dots, T-1;$
end for

شکل ۱: الگوریتم MSA پایه

- سرانجام بر روی منیفلد تعریف شده توسط معادله (۶) تصویرسازی می‌کنیم. این کار را با ماکسیمم‌سازی همیلتونین برای بدست آوردن $\boldsymbol{\theta^1} := \{\theta_t^1 : t = 0, \dots, T-1\}$ انجام می‌دهیم. یعنی برای t=0,..,T−1 قرار می‌دهیم $\theta_t^1 = \arg\max_{\theta\in\Theta_t} \sum_{s=1}^{S} H_t\left(x_{s,t}^{\boldsymbol{\theta^0}}, p_{s,t+1}^{\boldsymbol{\theta^0}}, \theta\right)$.

مراحل فوق آنقدر تکرار می‌شوند تا همگرایی حاصل شود. خلاصه الگوریتم MSA پایه در شکل ۱ آورده شده است.

۴. یک تقریب خطا برای MSA

با تمام مزیت‌های MSA، ممکن است الگوریتم فوق واگرا باشد مگر اینکه شرایط اولیه خوبی داده شده باشد. اجازه دهید ماهیت چنین پدیده‌هایی را با بدست آوردن یک تخمین خطای دقیق در هر تکرار از مراحل الگوریتم فوق بدست آوریم. ابتدا تعریف می‌کنیم:

$$W_t := \text{conv}\{x \in R^{d_t} : \exists \theta, s \quad \text{s.t.} \quad x_{s,t}^{\theta} = x\}$$

حال فرضیات زیر را در نظر بگیرید:

الف. Φ_s دوبار مشتق‌پذیر پیوسته است و Φ_s و $\nabla\Phi_s$ در شرط لیپ‌شیتز صدق می‌کنند. یعنی وجود دارد $K > 0$ به طوری که برای همه $x, x' \in W_t$ و $s \in [S]$ داریم:

$$|\Phi_s(x) - \Phi_s(x')| + \|\nabla\Phi_s(x) - \nabla\Phi_s(x')\| \le K\|x - x'\|$$

ب. $f_t(.,\theta)$ و $L_t(.,\theta)$ در x دوبار مشتق‌پذیر پیوسته هستند و f_t و $\nabla_x f_t$ و L_t و $\nabla_x L_t$ در شرایط لیپ‌شیتز پیوسته در x و لیپ‌شیتز یکنواخت در t و θ صدق می‌کنند یعنی وجود دارد $K > 0$ به طوری که برای همه $x, x' \in W_t$ و $\theta \in \Theta_t$ و t=1,…,T−1 داریم:

$$\|f_t(x,\theta) - f_t(x',\theta)\| + \|\nabla_x f_t(x,\theta) - \nabla_x f_t(x',\theta)\|_2 + |L_t(x,\theta) - L_t(x',\theta)|$$
$$+\|\nabla_x L_t(x,\theta) - \nabla_x L_t(x',\theta)\| \le K\|x - x'\|.$$

توجه کنید فرضیات الف و ب اگر هر W_t کران‌دار باشد که معمولا بوسیله کران‌داری Θ_t نتیجه می‌شود، به وضوح برقرارند. حال با استفاده از فرضیات فوق قضیه مهم زیر را مطرح می‌کنیم.

قضیه ۱.۴ (تخمین خطا برای MSA گسسته) فرض کنید فرضیات الف و ب برقرار باشند. آنگاه ثابت $C > 0$ مستقل از θ و ϕ وجود دارد به‌طوری که برای هر θ و $\Theta \in \Phi$داریم: $J(\phi) - J(\boldsymbol{\theta})$

$$\leq -\sum_{t=0}^{T-1}\sum_{s=1}^{S} H_t\left(x_{s,t}^{\boldsymbol{\theta}}, p_{s,t+1}^{\boldsymbol{\theta}}, \phi_t\right) - H_t\left(x_{s,t}^{\boldsymbol{\theta}}, p_{s,t+1}^{\boldsymbol{\theta}}, \theta_t\right) \qquad (7)$$

$$+\frac{C}{S}\sum_{t=0}^{T-1}\sum_{s=1}^{S}\left\|f_t\left(x_{s,t}^{\boldsymbol{\theta}}, \phi_t\right) - f_t\left(x_{s,t}^{\boldsymbol{\theta}}, \theta_t\right)\right\|^2 \qquad (8)$$

$$+\frac{C}{S}\sum_{t=0}^{T-1}\sum_{s=1}^{S}\left\|\nabla_x f_t\left(x_{s,t}^{\boldsymbol{\theta}}, \phi_t\right) - \nabla_x f_t\left(x_{s,t}^{\boldsymbol{\theta}}, \theta_t\right)\right\|^2 \qquad (9)$$

$$+\frac{C}{S}\sum_{t=0}^{T-1}\sum_{s=1}^{S}\left\|\nabla_x L_t\left(x_{s,t}^{\boldsymbol{\theta}}, \phi_t\right) - \nabla_x L_t\left(x_{s,t}^{\boldsymbol{\theta}}, \theta_t\right)\right\|^2 \qquad (10)$$

برهان. به [1] مراجعه شود.

برای تشریحی کوتاه از قضیه فوق، فرض کنید ϕ از $\boldsymbol{\theta}$ در گام بعدی الگوریتم گفته شده بدست آمده باشد در این صورت با توجه به گام ماکسیمم سازی تابع همیلتونین در الگوریتم واضح است که عبارت (۷) در قضیه فوق عبارتی نامثبت است. بنابراین با صرف‌نظر کردن از روابط ۱۰-۸ (این عبارات نامنفی را می‌توان بعنوان عبارات جریمه به تابع هدف اضافه کرد) داریم $J(\phi) - J(\boldsymbol{\theta}) \leq 0$ و این کاهش تابع هزینه را در هر گام نشان می‌دهد. برای تحلیل بیشتر قضیه به [1] مراجعه کنید.

۵. نتیجه‌گیری و کارهای مرتبط

در این مقاله از یادگیری عمیق با دیدگاه کنترل بهینه گسسته-زمان یاد شد و الگوریتمی کارا در حالت کلی برای این مسائل ارائه شد. در حالات خاص برای شبکه‌های عصبی باینری و سه‌قلو یعنی آن‌هایی که وزن‌ها در آن‌ها مجاز به انتخاب مقادیر به ترتیب از مجموعه‌های $\{-1,+1\}$ و $\{-1,0,+1\}$ هستند با توجه به الگوریتم پایه‌ای که در اینجا معرفی شد الگوریتم‌های متفاوت‌تر و مفیدتر معرفی می‌شود که به‌علت محدودیت قادر به بررسی آن‌ها در این متن نیستیم علاقه‌مندان می‌توانند برای بررسی بیشتر به [1] مراجعه کنند. در پایان خاطرنشان می‌شود که اکثریت و قسمت اصلی این متن برگرفته از [1] می‌باشد.

مراجع

1. Qianxiao Li, Shuji Hao, *An Optimal Approach to Deep Learning and Applications to Discrete-Weight Neural Networks* arXiv preprint arXiv:1803.01299, 2018.
2. Athans, M. and Falb, P. L. *Optimal control: an introduction to the theory and its applications*. Courier Corporation, 2013.
3. Chang, B., Meng, L., Haber, E., Ruthotto, L., Begert, D., and Holtham, E. *Reversible architectures for arbitrarily deep residual neural networks*. arXiv preprint arXiv:1709.03698, 2017.
4. Li, Q., Chen, L., Tai, C., and E, W. *Maximum principle based algorithms for deep learning. Journal of Machine Learning Research*, 18:1–29, 2018.