Science and AI Frontiers: Passing the Lower bound first: Writing on the Edging of Ambition and Greediness

Hengshuai Yao, email: hengshu1@ualberta.ca

December 26, 2023

Abstract

If you entered the field of scientific research like Artificial Intelligence research, the chances that you find a random paper with "This paper is the first time of proposing ...", "Our paper is the first ... method that ..."; ..., are really high. You will also find almost all the papers have this kind of sentences: "We call Equation * (or Algorithm *) the *WorldSuperbBestAlgorithm*". Naming something must feel you good. Isn't it? Have you cited similar algorithms? Have you cited the research that lead to your superb algorithm? Naming something without any reference or discussion is showing the world your paper is the first to discover the concept. I hope everyone understand this, and view naming something as a responsibility or a burden instead of pride. The motivation of this article is that AI authors should take caution in using these kinds of descriptions in promoting their works. It should always be backed up by extensive literature search, and the results of this search should be reflected in the paper, so that the readers have **materials and facts** to believe instead of merely consuming the authority of the authors.

Researchers often aim high, to pick apples that are high above. I think one should always pass the lower bound first. Writing your cherished paper close to or crossing the boundary of plagiarism isn't as hard as you might think. Don't claim you are the first in proposing "your ideas", unless you have done extensive literature search and 200% sure about what you are saying, not to mention intentionally hiding known arts to you just to establish and continue your leadership in the field.

Every paper is cooked in a hard way. After months and years of extensive efforts, when it is time to present the paper, one should always be aware to curb one's greediness in releasing your ambition in the paper. There is no need to laugh at this old-school ethics. Surprisingly, even famous people cannot guarantee doing this well. In my early career, I've experienced quite a few times being cut head in research.

1 My Problems

I have problems in reading AI papers. After reading so many papers that have "Ours is the first *** to ***", I am wondering, how many of these sentences are trustworthy? In fact, my bitter experience of Ph.D and early careers told me negatives. These words are solely used to catch the eyeballs of the reviewers and readers, and mean very little in concreteness.

I am a dim light in reinforcement learning and AI. However, I will use materials and facts to show that, I am proud that I am not doing this job a shabby way. **The point of this article is to raise awareness in the AI community about the trend of favoring taking the ownership of ideas and innovations**. Dr. Richard S. Sutton authored quite a few important papers in reinforcement learning, including Temporal difference learning, which was the first RL paper I read, linear Dyna, Gradient TD series, and the RL book which is used as a textbook in many universities.

Every Ph.D student has read about "How do you fail a Ph.D?". Below the materials are organized in "How one may fail to lead the field?". Never do "cutting-head" research. Never, never, never take advantage of your students. Don't do it the second time and third time, please, if you did it the first time already for some reasons that may be understandable. Never, never, never do evil.

2 Stay Out from "Cutting-head" research

Every research has a family tree, maintained by the scholars in literature. Respecting and building a genuine research tree is the shared honesty and ethics of the literature, which is the foundation of science and AI. "Cutting-head" research refers to, there exists relevant prior arts in the literature, in particular there is at least a paper A exists. However, the authors chose to present their paper "A0", with decades of research and writing experience, successfully wrapped A0 as the "first paper" in this literature, without any reference to paper A. Paper A then remains unknown to the literature, and later the literature cited and credited this idea to paper A0. Paper A's and the literature's head is cut.

This happens for the Gradient TD paper (Sutton et al., 2008a), written by Dr. Sutton in 2008. Gradient TD's key idea is to stabilize the O.D.E. of TD update, which is problematic for off-policy learning. The paper presents GTD was "ground breaking", in the way that the O.D.E. update was conceived and the learning objective and the gradient descent for TD were novel. I started writing the preconditioned TD paper (Yao and Liu, 2008) from 2006, and submitted the paper four times to ICML 2007, NIPS 2007 (see my homepage, a rejection decision with 7.27 average score out of 10, three reviewer scores: 7,8 and 5, confidence of the reviewers: 10, 10, 2. The third reviewer gave 2 in the first round of review.), ISAIM 2008, and ICML 2008. Finally, the paper got accepted by ICML 2008. Dr. Sutton's paper is NIPS 2008, which is half a year later than ICML 2008.

What does GTD do? It is the first off-policy temporal difference (TD) learning that is convergent with O(n) complexity. Its contributions are three things according to the paper:

- "The gradient temporal-difference (GTD) algorithm estimates the expected update vector of the TD(0) algorithm and performs stochastic gradient descent on its L2 norm". (GTD abstract)

3 The GTD(0) algorithm

We next present the idea and gradient-descent derivation leading to the GTD(0) algorithm. As discussed above, the vector $\mathbb{E}[\delta\phi]$ can be viewed as an error in the current solution θ. The vector should be zero, so its norm is a measure of how far we are away from the TD solution. A distinctive feature of our gradient-descent analysis of temporal difference learning is that we use as our objective function the L_2 norm of this vector:

$$J(\theta) = \mathbb{E}[\delta\phi]^\top \mathbb{E}[\delta\phi]\,. \tag{5}$$

This objective function is quadratic and unimodal; it's minimum value of 0 is achieved when $\mathbb{E}[\delta\phi] = 0$, which can always be achieved. The gradient of this objective function is

$$\begin{aligned}\nabla_\theta J(\theta) &= 2(\nabla_\theta \mathbb{E}[\delta\phi])\mathbb{E}[\delta\phi] \\ &= 2\mathbb{E}\big[\phi(\nabla_\theta\delta)^\top\big]^\top \mathbb{E}[\delta\phi] \\ &= -2\mathbb{E}\big[\phi(\phi-\gamma\phi')^\top\big]^\top \mathbb{E}[\delta\phi]\,. \end{aligned} \tag{6}$$

This last equation is key to our analysis. We would like to take a stochastic gradient-descent approach, in which a small change is made on each sample in such a way that the expected update

A natural idea is that the current weights can be improved by minimizing the residual error $||e_{t+1}(w)||^2$, which produces a gradient descent algorithm

$$w_{t+1} = w_t - \alpha_t A'_{t+1}(A_{t+1}w_t + b_{t+1}),$$

where α_t is a positive step-size. Gradient descent algorithm is a stochastic form of the iteration (5).

The general preconditioned temporal difference (PTD) learning applies the technique of preconditioning to improve the convergence rate of gradient descent. Assume C_{t+1} is a chosen preconditioner, the rule of PTD can be cast as

$$w_{t+1} = w_t - \alpha_t C_{t+1}^{-1} A'_{t+1}(A_{t+1}w_t + b_{t+1}), \tag{14}$$

Figure 1: Left: from GTD paper. Right: from the preconditioning paper.

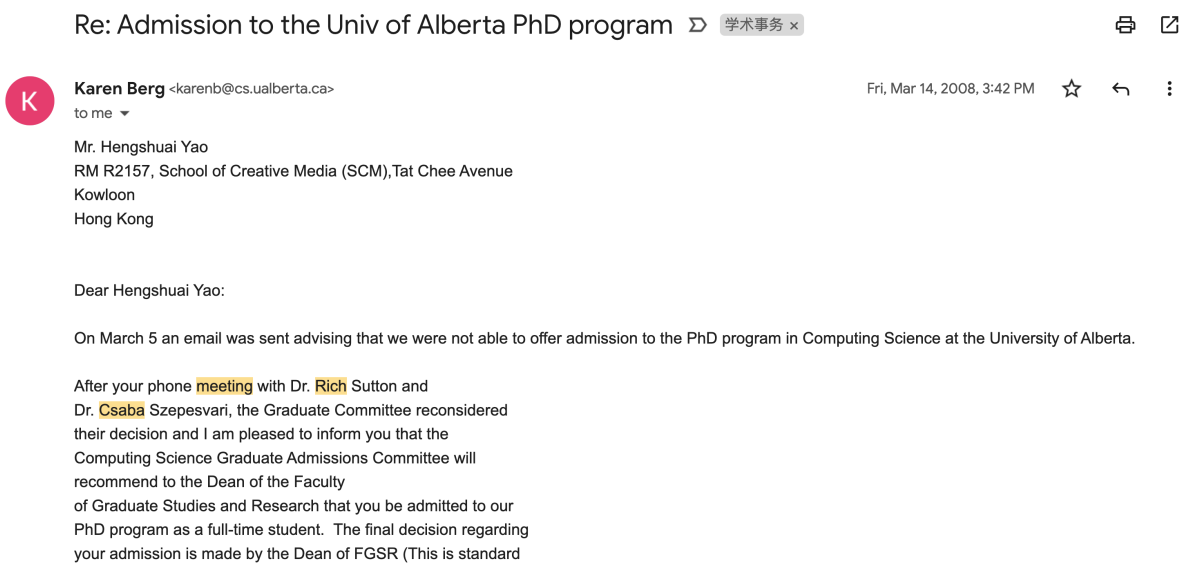
Prior to the current work, the possibility of instability could not be avoided whenever four individually desirable algorithmic features were combined: 1) off-policy updates, 2) temporal-difference learning, 3) linear function approximation, and 4) linear complexity in memory and per-time-step computation. If any one of these four is abandoned, then stable methods can be obtained relatively easily. But each feature brings value and practitioners are loath to give any of them up, as we discuss later in a penultimate related-work section. In this paper we present the first algorithm to achieve all four desirable features and be stable and convergent for all finite Markov decision processes, all target and behavior policies, and all feature representations for the linear approximator. Moreover, our algorithm does not use importance sampling and can be expected to be much better conditioned and of lower variance than importance sampling methods. Our algorithm can be viewed as performing stochastic gradient-descent in a novel objective function whose optimum is the least-squares TD solution. Our algorithm is also incremental and suitable for online use just as are simple temporal-difference learning algorithms such as Q-learning and TD(λ) (Sutton 1988). Our algorithm can be broadly characterized as a gradient-descent version of TD(0), and accordingly we call it GTD(0).

Figure 1: GTD Introduction

I understand that the GTD algorithm applies to off-policy while my PTD paper does on-policy learning only. In addition, GTD is O(n) and my PTD is $O(n^2)$. However, (1) the GTD paper also tried to cover on-policy. (2) The Gradient Descent for TD idea: is it first in the GTD paper or my PTD paper? (3) When were this objective function and the O.D.E. first time appearing in literature? in GTD paper or my PTD paper? (4) How hard is it to conduct the GTD analysis once you got the idea of the symmetry in the system from the objective function? (5) Without this idea, how would one stabilize TD for off-policy learning? Any clue for this before 2008?

3 Don't apply any Filter. Science is Objectivity.

I applied the Ph.D program at University of Alberta, in December 2007. I uploaded my CV, research statement and my PTD paper (my only paper at the time). Dr. Sutton interviewed me in March, 2008, in which we discussed my research on TD and PTD. Karen's email shows this interview in Figure 2.



Re: Admission to the Univ of Alberta PhD program 学术事务 ×

Karen Berg <karenb@cs.ualberta.ca> Fri, Mar 14, 2008, 3:42 PM
to me

Mr. Hengshuai Yao
RM R2157, School of Creative Media (SCM),Tat Chee Avenue
Kowloon
Hong Kong

Dear Hengshuai Yao:

On March 5 an email was sent advising that we were not able to offer admission to the PhD program in Computing Science at the University of Alberta.

After your phone meeting with Dr. Rich Sutton and
Dr. Csaba Szepesvari, the Graduate Committee reconsidered
their decision and I am pleased to inform you that the
Computing Science Graduate Admissions Committee will
recommend to the Dean of the Faculty
of Graduate Studies and Research that you be admitted to our
PhD program as a full-time student. The final decision regarding
your admission is made by the Dean of FGSR (This is standard

Figure 2: Karen's email about the interview.

A search of my email box showed lots of communications between Dr. Sutton's group at the time and me, before their NIPS 2008 submission, including,

- Email chains with Dr. Sutton himself. Figure 3.
- My experiments for GTD on Boyan chain. Figure 4.
- My PDF two-page write-up of GTD experiments, pointing out two problems of GTD. Figure 5.
- Dr. Sutton asked me for the matlab code of my GTD experiments. Figure 9.
- My email attachments to Dr. Sutton, including codes, write-up PDF, plots and relevant papers. Figure 6.

I had a hard time reading the Acknowledgement of GTD. Retracing this route, it gave me an impression that Dr. Sutton had a filter in writing down acknowledgement, [1] and "chose" which students to work on what, regardless how much interests and contributions one student already made to the project. I believe all these people contributed to your paper's discussion. However, as someone who conceived some experiments, wrote codes, wrote a two-page PDF and had lots of email discussion with the authors, when one wrote down the names to acknowledge, the student don't even came across into the writer's mind? What is the purpose of the Acknowledgement in a paper? It shows the authors' appreciation for them spending time giving feedback, so that people know you appreciate. It is always

[1]Did I over-read? I read almost all Dr. Sutton's writings on off-policy learning, including his new version of the book, ETD, GQ and nonlinear GTD; etc. I hope I could have found some clue that "He just forgot".

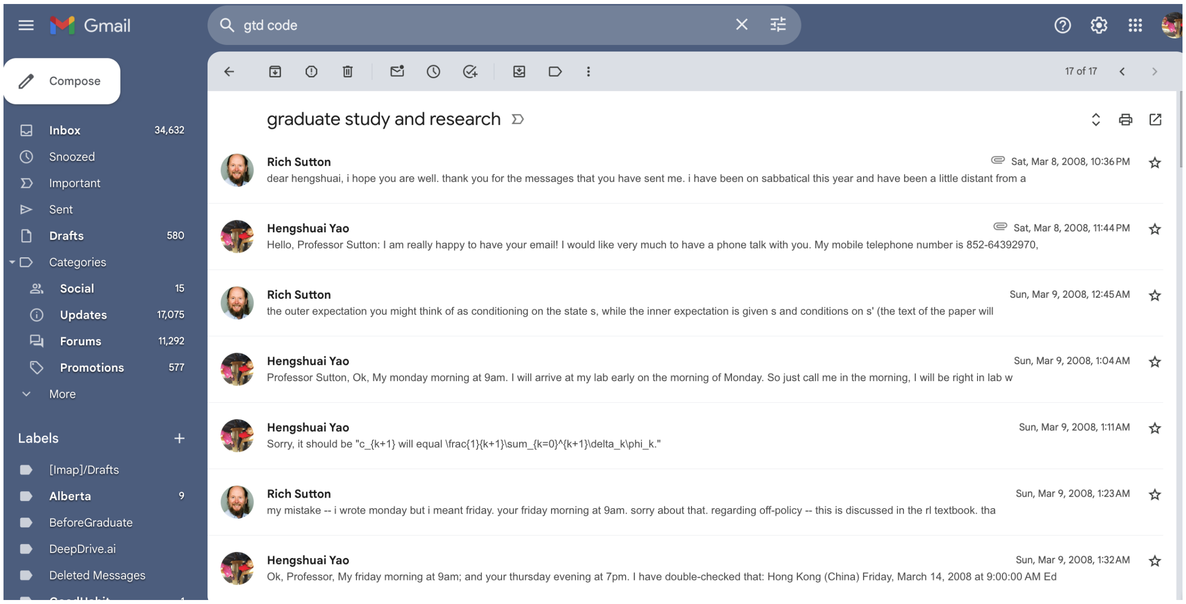


Figure 3: Emails with Dr. Sutton and me.

good to genuinely and generously thank the people who gave feedbacks, even though their opinions are not very important to the paper itself now (otherwise they should be put as the coauthors).

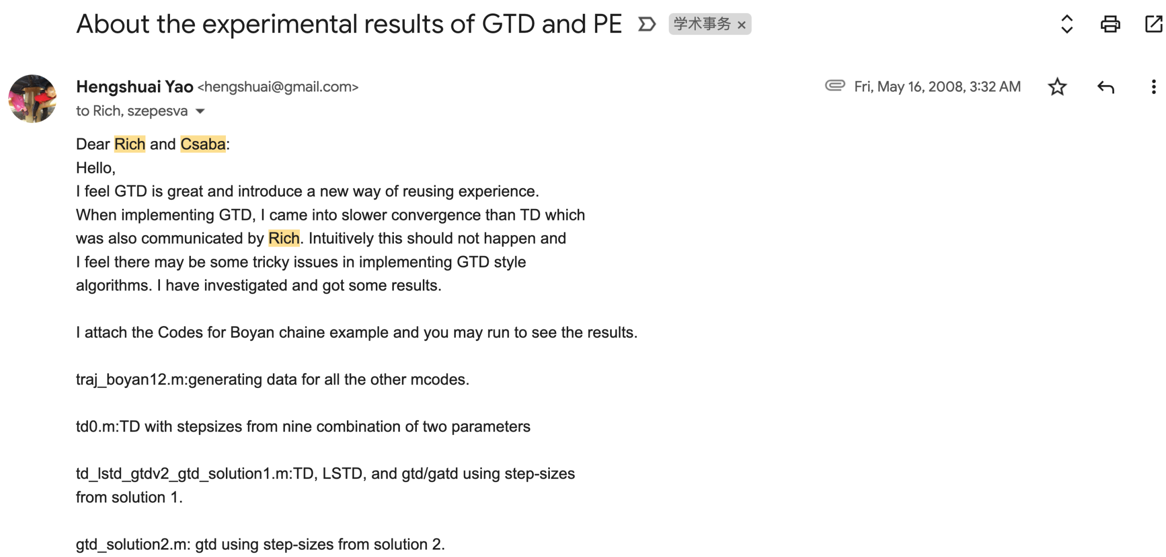
LET US THROW AWAY OUR SUBJECTIVE FILTERS. "This author(s) did not cite my paper. So I will not cite hers/his as well". "This author spoke bad of my papers and so I hate her/him; and of course I will not cite her/his papers". Believe me, if you curb all these mental things, you will feel much more happy than if you submerge in these thoughts. Cite objectively. One is one. Two is Two. The research sky needs to be clear and blue.

4 Don't do it the 2nd time, please.

I came to UofA September 2008. About the same time, the development of TDC paper started (Sutton et al., 2009). My involvement can be reflected in

- Figure 12 shows I help organize this meeting from the beginning. These offline meetings were scheduled weekly, which lead to the TDC paper.
- Figure 13 shows the idea of TDC started with Hamid relating to the preconditioning technique from my PTD paper. Hamid and I had lots of offline discussions as we sat in the lab everyday at the time.
- Emails from David Silver Figure 7 and 8 on TDC experiments.

Yet, you won't be able find any mentioning of my above involvements in GTD or TDC papers in official documents. GTD missed my name even in the Acknowledgement. TDC magically doesn't have the Acknowledgement section. The TDC paper started with the weekly meetings, which were initiated and scheduled from the discussion between Dr. Sutton and me, Figure 12. Boyan chain experiments? Who did it? Where the code was from? All the experiments used on-policy problems for evaluating off-policy algorithms. Whose idea? Figure 5? This also includes Silver's Computer Go experiments for comparing with TD. Is off-policy learning algorithm comparable to on-policy learning algorithm? Whose idea? Silver's communication was addressing me first, Figure 7 and Figure 8. I discovered this simple yet novel way of evaluating GTD and other off-policy learning algorithms at the time when I was a member of the "Off-policy Gangs". Yet, I have no credit in the paper. I have no credit in this now widely used technique for off-policy learning experiments, which was invented by me. I've seen too many papers like TDRC (Ghiassian et al., 2020) and a few others by actually some friends and



About the experimental results of GTD and PE 学术事务 ×

Hengshuai Yao <hengshuai@gmail.com> Fri, May 16, 2008, 3:32 AM
to Rich, szepesva

Dear Rich and Csaba:
Hello,
I feel GTD is great and introduce a new way of reusing experience.
When implementing GTD, I came into slower convergence than TD which
was also communicated by Rich. Intuitively this should not happen and
I feel there may be some tricky issues in implementing GTD style
algorithms. I have investigated and got some results.

I attach the Codes for Boyan chaine example and you may run to see the results.

traj_boyan12.m:generating data for all the other mcodes.

td0.m:TD with stepsizes from nine combination of two parameters

td_lstd_gtdv2_gtd_solution1.m:TD, LSTD, and gtd/gatd using step-sizes
from solution 1.

gtd_solution2.m: gtd using step-sizes from solution 2.

Figure 4: Email showing I did experiments for GTD.

colleagues referring the technique, the NEU objective function, and the gradient TD idea to the GTD and TDC papers. However, I cannot tell them. It's a bit pain to read sometimes.

Chapter 11 of Dr. Sutton's book, a whole chapter, including historical contributions, you won't be able to see any mentioning of my work and contributions either.

5 First-authoring as Supervisor: Caution and Use Rarely

I can only understand Dr. Sutton's behavior in terms of his ambition. I highly appreciate his ambition and vision for RL. However, in one's pursuit of ambition for himself and literature, don't forget to curb our human instinctive feelings.

Does one need to be the first author for the papers coauthored with students? I absolutely agree, the supervisors can, definitely, in some situations. **Actually, it is great to see some senior folks proposing new ideas and take a big chunk of the work in the frontiers of research**. For Dr. Sutton, this happened a bit too often. GTD (Sutton et al., 2008a), TDC (Sutton et al., 2009), Horde (Sutton et al., 2011), Linear Dyna (Sutton et al., 2008b) and ETD (Sutton et al., 2016); etc. Most people are Okay with it, especially young students and scientists. However, it is a completely different story, if one left some coauthor(s) out, and if one is the first author and correspondence author of the paper. I would be Okay if I was put the last author of TDC, although reading the paper, and eyeing the emails and trying to recall by *inaccurate memories*, I *may* qualify as the first four authors. It is *meaningless* to compare contributions. I am just saying. You get my point.

A reminder to all of us: Don't forget to search your email boxes, memory, and your coauthors' minds about potential coauthors when one submits. This is especially important because nowadays big AI conferences does not allow the changes of authors once submitted.

If one really loves your own ideas so much, be advised that you can always write a paper on your own. One of the toughest times I've been through is that the GTD matter and other mentioned involved my genuine some life-long friends as well. Don't involve coauthors if there are any unknown substantial facts to them. The coauthors and colleagues join in a work because they want to help in making a great paper, and they can make it better. They are NOT supposed to be kept in any hideous intention, or take blame in the future for which they didn't know well or didn't do.

The content of the following papers used to inspire me. However, the author line doesn't look so right to me now.

- A Convergent O(n) Algorithm for Off-policy Temporal-difference Learning with Linear Function Approximation. Richard S. Sutton, Hamid Maei and Csaba Szepesvári. NeurIPS, 2008.

- Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation. Richard S. Sutton, Hamid Reza Maei, Doina Precup Shalabh Bhatnagar, David Silver, Csaba Szepesvari, and Eric Wiewiora. ICML, 2009.
- Horde: A Scalable Real-time Architecture for Learning Knowledge from Unsupervised Sensorimotor Interaction. Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White. AAMAS, 2011.
- An Emphatic Approach to the Problem of Off-policy Temporal-Difference Learning. Richard S. Sutton, A. Rupam Mahmood and Martha White. JMLR, 2016.

6 Afterword

I've chosen to forget. During the 15 years, I fighted with myself on and on. Every time I succeeded. This effort collapsed when reading recent off-policy RL papers, one after another, including some written by Dr. Sutton himself. I give up. The amount of pain grows beyond that I can continue to withhold any more.

I understand this may cost my research career. This document is my own decision. No one else and no company was involved in the discussed matters.

The content of this article reflects what I thought and read over the years. It is a subjective matter that is my own thoughts. I apologize if it is disturbing to your mind.

I believe, **matters like this, should be discussed with the people in question before releasing to public**. Please give people chances to chip in, if you have similar cases. They may want to explain or make up their mistakes. Everyone makes mistake, including me. I did quite a few homeworks before this. Folks at UofA, including some coauthors of GTD and TDC, senior members of Amii, department chairs, vice dean and dean of faculty of science, and provost, all witnessed my efforts and patience to try to resolve this privately, from March 2022.

I think by the time this article is published (if I decide after all), the time is sufficient enough for someone to respond responsibly. Dr. Sutton ignored my personal communications for about half year. After that, I contacted the dean, who confidently assured me to let him do the communication efforts to Dr. Sutton and he sounded very hopeful to help resolve it. I was very hopeful too. Dr. Sutton met me November 3rd, 2022, in his office. We discussed the TDC paper and my PTD paper in particular. He recognized the similarity by himself after he pulled the two papers, and said TDC is a "warping" of PTD. I didn't know this word before. He gave me a "sorry" which sounded quite light, and he refused to do any fix. I especially didn't like it when he said, "it's such a long time.", which sounded like why I bothered to mention it now. I don't know how to describe my feelings. For my 15 years. For my young and passionate times on Gradient TD works. For my considerations for him. I get this.

I said, "I really need to have break now". I still shook my hands with Dr. Sutton. When I walked out of his office, I saw Rupam, an old lab mate at our times, and I said "Hi". Since that meeting, I never heard from him.

7 Declarations

7.1 Ethics approval and consent to participate Consent for publication

YES.

7.2 Availability of data and material Competing interests

No data and no competing interests.

7.3 Funding

No funding for this article.

7.4 Authors' contributions

Me only.

7.5 Acknowledgements

No acknowledgement is necessary for this article.

Appendix: Timeline

- From 2006 to 2008, I have submitted my PTD work four times. I was a master student at Tsinghua Univ. and CityU at Hong Kong at the time.
- In Dec. 2007, I applied to UofA Ph.D program, with my PTD paper in the application material.
- January 2008: my PTD paper was finally accepted by ICML.
- From March to May 2008, extensive discussions between Dr. Sutton's group and me. I also did experiments, and one theorem proof.
- May 2008, Dr. Sutton's GTD paper was submitted. No my name. either in the author or Acknowledgement.
- September 2008, I came to UofA, under the supervision of Dr. Sutton.
- About the same time, TDC paper started to be developed. My involvement is shown in this article.
- January 2009, TDC paper was submitted. Again. No my name. The paper this time has no Acknowledgement section.
- From 2009 to 2011, I was exploring multi-step Dyna, off-policy learning with a one-collection-for-all solution, Universal option models, and RL for PageRank.
- Dr. Sutton's Horde, the concurrent work of my one-collection-for-all off-policy learning solution, and his ETD both have shadows in my papers.
- Dr. Sutton told me in 2011 that he would not be my supervisor any more. FGSR at UofA and quite a few professors knew this history. I became a new father before or right after this, and my wife was visiting me and she did not have a job. I held an international visa at the time.
- Luckily, Csaba took me as a student later on.
- In 2014, Csaba organized my defense committee. He told me that he asked Dr. Sutton whether he could be a member. Dr. Sutton said "NO". I didn't ask Csaba to ask Dr. Sutton. From his point of view, it must have been natural: I worked with Dr. Sutton for three years, and my thesis is RL. One can read the Acknowledgement of my Ph.D thesis (https://github.com/hengshuaiyao/HengshuaiYao.github.io/blob/master/papers/yao_hengshuai_PhD.pdf) and see words I had for Dr. Sutton.
- January 2015: End of my student-hood at UofA.
- After 2015: Dr. Sutton treated me a bit better than a stranger.

Obviously my academic career will not be able to continue any more. Continuing to hold this over the years has hurt my mental health and my family life. I need to give my mind a break. I need to give my family a happy Christmas with no clouds on a cheerful Daddy's mind.

References

Ghiassian, S., Patterson, A., Garg, S., Gupta, D., White, A., and White, M. (2020). Gradient temporal difference learning with regularized corrections. In *Proceedings of the 37th International Conference on International Conference on Machine Learning*.

Sutton, R. S., Maei, H., and Szepesvári, C. (2008a). A convergent $O(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in neural information processing systems*, 21.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, pages 993–1000.

Sutton, R. S., Mahmood, A. R., and White, M. (2016). An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1):2603–2631.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., and Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 761–768.

Sutton, R. S., Szepesvári, C., Geramifard, A., and Bowling, M. (2008b). Dyna-style planning with linear function approximation and prioritized sweeping. UAI'08, page 528–536, Arlington, Virginia, USA. AUAI Press.

Yao, H. (2023). A new Gradient TD algorithm with only one step-size: Convergence rate analysis using l-λ smoothness. *arXiv:2307.15892*.

Yao, H. and Liu, Z.-Q. (2008). Preconditioned temporal difference learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1208–1215, New York, NY, USA. Association for Computing Machinery.

Experimental results of on-policy GTD and one variant

Abstract

I investigated and found there are two significant issues that are the key to the performance of GTD. The first is that GTD is different from TD in that it is two-time scale, and there are two learning parameters using their independent step-size. This makes the tuning of algorithm very hard because the two step-sizes have a mutual effect on each other. The second is that there are noises in GTD that influence the convergence rate significantly. I used the update direction of iLSTD to help decrease the effects of noises and developed an adaptive step-size for vector c. Results of on-policy GTD on Boyan chain example is encouraging.

1 A variant of GTD: Gradient-Averager TD

A variant of GTD is to directly use the averaged gradient information of TD as the update direction, producing

$$\theta_{k+1} = \theta_k + \alpha_k c_{k+1}, \tag{1}$$

where α_k is a positive step-size. We call it Gradient-Averager TD(GATD).

2 Examining the tricky issues in implementing GTD and GATD

The update for c is

$$c_{k+1} = c_k + \beta_k(\delta_k \phi_k - c_k), \tag{2}$$

where δ_k is the TD signal and ϕ_k is the feature of current state. Notice that vector c is supposed to provide the direction of iLSTD. However, some noise is introduced here because the direction of iLSTD is to combine all the experience with the latest weights:

$$\bar{c}_{k+1} = \frac{1}{k+1}\sum_{j=0}^{k} \phi_j(\gamma\phi'_j - \phi_j)^T\theta_{k+1} + \frac{1}{k+1}\sum_{j=0}^{k} \phi_j r_j = A_k\theta_k + b_{k+1}.$$

Figure 5: My write-up for the GTD experiments. This write-up was in *May 2008*, and sent to Dr. Sutton in one of the email attachments. The document shows that I pointed out two problems of implementing GTD, which remained unsolved until my recent Impression GTD (Yao, 2023). It also showed I started experimenting using averaging to curb the noises in GTD, which were proposed in a few recent papers, including NASA and some iterative averaging work (See my Impression GTD for the review).

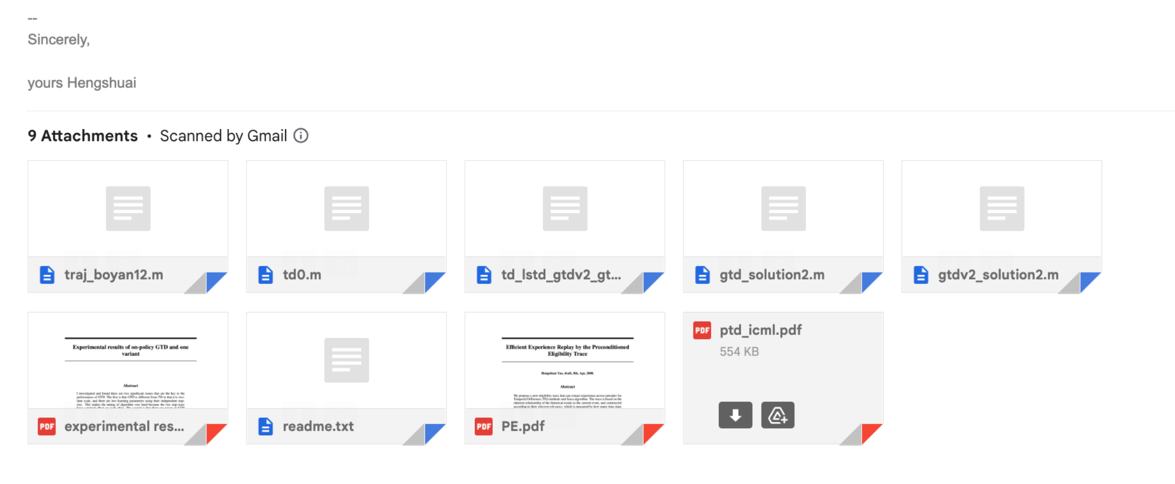
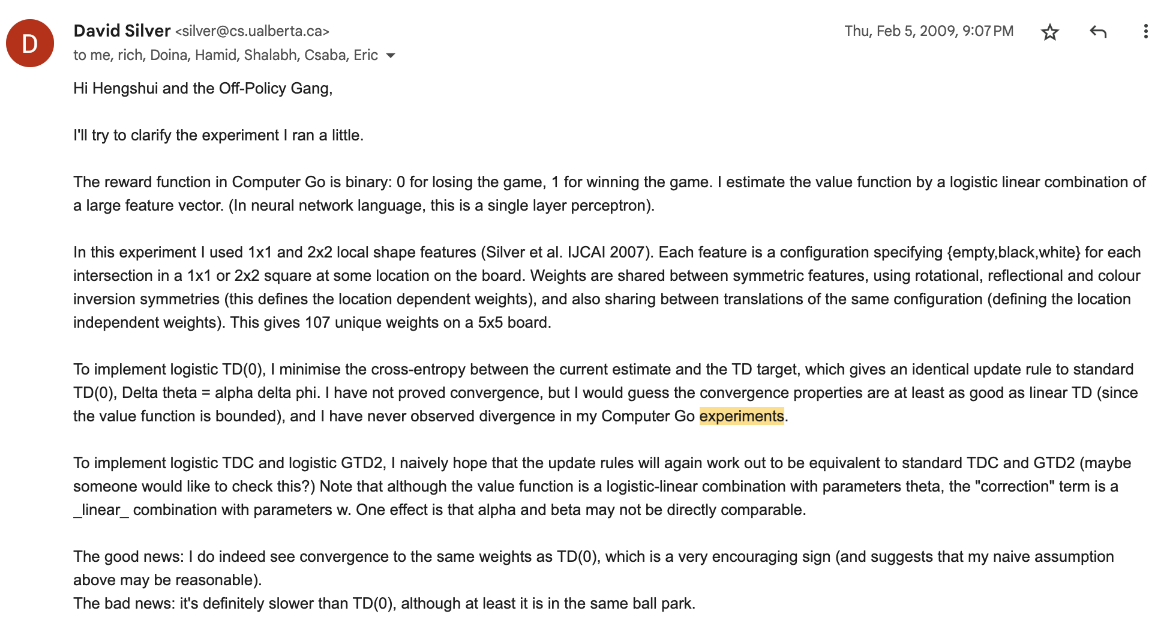


Figure 6: Email attachments by me in May 2008.



David Silver <silver@cs.ualberta.ca> Thu, Feb 5, 2009, 9:07 PM

to me, rich, Doina, Hamid, Shalabh, Csaba, Eric

Hi Hengshui and the Off-Policy Gang,

I'll try to clarify the experiment I ran a little.

The reward function in Computer Go is binary: 0 for losing the game, 1 for winning the game. I estimate the value function by a logistic linear combination of a large feature vector. (In neural network language, this is a single layer perceptron).

In this experiment I used 1x1 and 2x2 local shape features (Silver et al. IJCAI 2007). Each feature is a configuration specifying {empty,black,white} for each intersection in a 1x1 or 2x2 square at some location on the board. Weights are shared between symmetric features, using rotational, reflectional and colour inversion symmetries (this defines the location dependent weights), and also sharing between translations of the same configuration (defining the location independent weights). This gives 107 unique weights on a 5x5 board.

To implement logistic TD(0), I minimise the cross-entropy between the current estimate and the TD target, which gives an identical update rule to standard TD(0), Delta theta = alpha delta phi. I have not proved convergence, but I would guess the convergence properties are at least as good as linear TD (since the value function is bounded), and I have never observed divergence in my Computer Go experiments.

To implement logistic TDC and logistic GTD2, I naively hope that the update rules will again work out to be equivalent to standard TDC and GTD2 (maybe someone would like to check this?) Note that although the value function is a logistic-linear combination with parameters theta, the "correction" term is a _linear_ combination with parameters w. One effect is that alpha and beta may not be directly comparable.

The good news: I do indeed see convergence to the same weights as TD(0), which is a very encouraging sign (and suggests that my naive assumption above may be reasonable).
The bad news: it's definitely slower than TD(0), although at least it is in the same ball park.

Figure 7: David Silver emails.

The good news: I do indeed see convergence to the same weights as TD(0), which is a very encouraging sign (and suggests that my naive assumption above may be reasonable).
The bad news: it's definitely slower than TD(0), although at least it is in the same ball park.

See also my comments below. Hope this helps!
-Dave

On 5-Feb-09, at 7:19 PM, Hengshuai Yao wrote:

> Hi, David, could you explain a little bit more about the features?
> Although Rich told me already, but I'd like to make sure are you using nonlinear Neural networks (3 layed? sigmoidal?), not linear function approximation?
>
> Does TD still converge for this nonlinear function approximation? That's amazing.
>
> On Thu, Feb 5, 2009 at 1:38 PM, Rich Sutton <rich@richsutton.com> wrote:
>
>> fyi. perhaps we are not done yet. perhaps we are back almost to where we were at nips, except now we know where we are a little better.
>> r
>>
>> Begin forwarded message:
>>
>> **From:** David Silver <silver@cs.ualberta.ca>
>> **Date:** February 3, 2009 4:40:36 PM MST (CA)
>> **To:** Rich Sutton <sutton@cs.ualberta.ca>

Figure 8: David Silver emails.

Rich Sutton <rich@richsutton.com>
to me, Csaba
May 17, 2008, 2:59 PM

yes. your writeup has many interesting ideas. i can start with some superficial comments. the initial averaging idea (1) seems similar to the momentum idea from neural networks. am i right in thinking the gatd line of thought does not really say anything about converging off-policy? (that's what you meant by no lyapunov fn?) in section 3, did you slip in an alternate notation 'w' for the weights -- is this the same as theta or something different?

would you mind if we shared your work on this with another student who has been working with us on gtd? hamid reza-maei would find it all very interesting and is also a matlab user.

rich

...

...
<experimental results.pdf><boyan_N_small.eps><boyan_N_small.ps><solution1.eps><solution1.ps><solution2.eps><solution2.ps><experimental results.tex><error.eps>
<error.ps>

Hengshuai Yao <hengshuai@gmail.com>
to Rich, Csaba
May 17, 2008, 7:49 PM

On 5/18/08, **Rich Sutton** <rich@richsutton.com> wrote:

> yes. your writeup has many interesting ideas. i can start with some superficial comments. the initial averaging idea (1) seems similar to the momentum idea from neural networks. am i right in thinking the gatd line of thought does not really say anything about converging off-policy? (that's what you meant by no lyapunov fn?) in section 3, did you slip in an alternate notation 'w' for the weights -- is this the same as theta or something different?

Thanks!

-f:1269818092513078964&th=119f4cd3cd9b2eb4&view=att&disp=inline&realattid=f_fgbl4nbe

Figure 9: Dr. Sutton asked my matlab code for the GTD experiments.

Hamid Reza Maei <maei@cs.ualberta.ca>
May 18, 2008, 10:47 PM

to me, Rich, Csaba

Hi Hengshuai,

I went through your communication with Rich and Csaba with interest and also your icml paper. It was very interesting and encouraging. We should certainly keep continuing these discussions. I have a few questions or maybe comments:

1. In gtd(0) algorithm, given the fact that c parameter is known, how did/would you compare the speed of convergence for gtd(0) and td(0)?
I have tried that for Random walk experiment and I found gtd(0) still is slower than td(0) specially when the number of states increases. If that is the case for Boyan chain, then we may conclude that gtd(0) in general is slower than td(0), thus tuning the learning rate parameter, beta (the one that controls c), won't help to increase the speed of convergence.

2. (this is rather comment) When I look at gatd method, I don't see how it might help the convergence for off-policy learning. Following Tsitsiklis & Van Roy line of proof,
the asymptotical result will be the same as TD solution, A*theta+b=0, and if A is not negative definite, then
the algorithm won't converge. The only thing that I could think of is that using gatd method make the convergence somehow faster such as batch learning algorithm--if I am right, with less number of data but it generally requires the model of environment, if I understood the model correctly from your write up. Also shouldn't the asymptotical result of gatd(0) and td(0) be very similar? Is you RMS averaged over number of trials as well?
I also thought that by increasing the number of experiment trials (runs), the RMS result for td(0) and gatd(0) should be pretty close given large number of trials. Let me know what you think.

Thanks:)
Hamid
On 17-May-08, at 11:43 PM, Rich Sutton wrote:

Figure 10: Hamid emails.

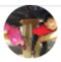
absolute abilities not previous available in existing algorithms. We have conducted empirical studies with the GTD(0) algorithm and have confirmed that it converges reliably on standard off-policy counterexamples such as Baird's (1995) "star" problem. On on-policy problems such as the n-state random walk (Sutton 1988; Sutton & Barto 1998), GTD(0) does not seem to learn as efficiently as classic TD(0), although we are still exploring different ways of setting the step-size parameters, and other variations on the algorithm. It is not clear that the GTD(0) algorithm in its current form will be a fully satisfactory solution to the off-policy learning problem, but it is clear that is breaks new ground and achieves important abilities that were previously unattainable.

Acknowledgments

!

The authors gratefully acknowledge insights and assistance they have received from David Silver, Eric Wiewiora, Mark Ring, Michael Bowling, and Alborz Geramifard. This research was supported by iCORE, NSERC and the Alberta Ingenuity Fund.

Figure 11: GTD Conclusion and Acknowledgement



Hengshuai Yao <hengshuai@gmail.com>
to Rich

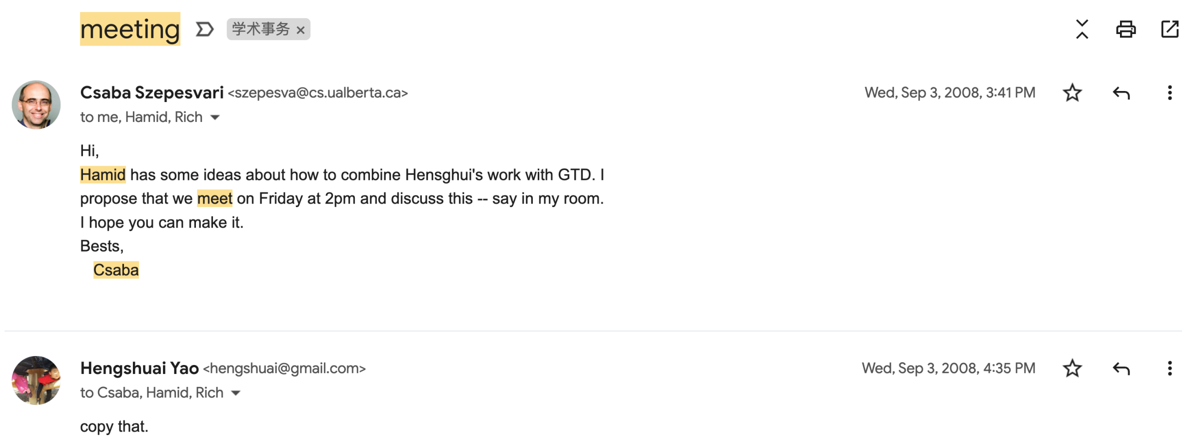
Could we figure out another time, as I have a course on 3PM, Wen.
I have course on
3PM--4:30, Mon, Wen
11:00AM--12:20, Tue, Thur.
The other time is OK.

---------- Forwarded message ----------
From: Csaba Szepesvari <szepesva@cs.ualberta.ca>
Date: Mon, Oct 27, 2008 at 12:48 PM
Subject: Re: Hi, off-policy TD meeting at Wendenesday
To: Hamid Reza Maei <maei@cs.ualberta.ca>
Cc: Eric Wiewiora <wiewiora@cs.ualberta.ca>, Hengshuai Yao
<hengshuai@gmail.com>, shalabh@csa.iisc.ernet.in, Rich Sutton
<rich@richsutton.com>

Yes

Hamid Reza Maei wrote:
>
> That sounds good. Wed. at 2PM also sounds good to me. What about Csaba? Can he make it?
>
> Cheers,
> Hamid
>
> On 27-Oct-08, at 12:49 PM, Eric Wiewiora wrote:
>
>> Wednesday at 2 is fin for me.
>>
>> - Eric
>>
>> On Oct 27, 2008, at 12:46 PM, Hengshuai Yao wrote:
>>
>>> Hi, all, Rich would like to organize an off-policy TD meeting at 2PM,
>>> Wen., his office.
>>>
>>> Please tell me if you could not make it, and suggest your time slots.
>>> Also Rich suggests make it a Regular "off-policy " Meeting every week.
>>>
>>> Thanks!
>>> --
>>> Sincerely,
>>>
>>> yours Hengshuai
>>

Figure 12: Offline meetings. It looks I helped organize these meetings.



meeting 学术事务 ×

Csaba Szepesvari <szepesva@cs.ualberta.ca> Wed, Sep 3, 2008, 3:41 PM
to me, Hamid, Rich

Hi,
Hamid has some ideas about how to combine Hensghui's work with GTD. I
propose that we meet on Friday at 2pm and discuss this -- say in my room.
I hope you can make it.
Bests,
Csaba

Hengshuai Yao <hengshuai@gmail.com> Wed, Sep 3, 2008, 4:35 PM
to Csaba, Hamid, Rich

copy that.

Figure 13: TDC Idea source.