On the GTD paper and TDC/GTD2 paper

Hengshuai Yao
Huawei Technologies Canada,
University of Alberta,
Edmonton, AB, Canada

July 15, 2022

Abstract

In this note, I discuss the technical and ethical aspects of the GTD paper and the TDC paper, both by Rich S. Sutton's group to which I also used to belong at the time. The intention of the note is that both papers failed to cite my preconditioning paper appropriately. Rich was aware of my paper at the time. The discussion below is not intended to upset anyone, and the main body is mostly focused on technical matters.

1 Papers in Question

There are three papers in question in this note:

- GTD paper: A Convergent O(n) Algorithm for Off-policy Temporal-difference Learning with Linear Function Approximation, Richard S. Sutton, Hamid Maei, Csaba Szepesvari. NIPS, 2008.
- TDC paper: 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 Szepesvári, Eric Wiewiora. ICML, 2009.
- Preconditioning paper: Preconditioned Temporal Differnce Learning. Hengshuai Yao and Zhi-Qiang Liu, ICML, 2008.

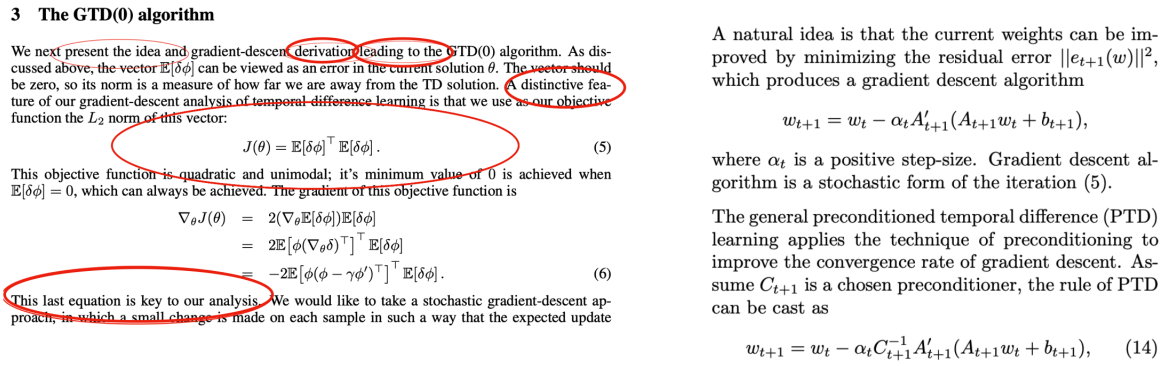
2 GTD paper

What does GTD do? It is the first off-policy temporal difference (TD) learning that is convergent with O(n) complexity. Its contribution 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)
- O(n) complexity. This is highlighted in the title (but not mentioned in the abstract).
- An analysis of the convergence.

In this note, I point out that the first item is not their contribution. In the preconditioning paper, this idea was proposed already. That is, the L2 objective, and the gradient descent minimizing "the expected update vector of the TD(0) algorithm" is the contribution of preconditioning paper. However, the GTD paper didn't acknowledge this important source of idea when they derived their algorithm from this objective, nor did they mention it across the whole paper, nor was I even acknowledged in both GTD and TDC papers.

I also participated a few meetings of the TDC paper. Rich said "this time you can be an author as well". However, my authorship was removed after a technical argument with one of the authors. It wasn't like a general group meeting. I was the only one who participated in the weekly technical



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.

meetings on this paper but not on the final author list. No one explained to me any reason for that. Nor did I talk to anyone including Rich about this at the time. I thought that's Okay. This paragraph is all dependent on my memories which may not be reliable after so many years. Below I will cite simple facts from the papers involved.

Figure 1 compares the most obvious parts from both papers. A reinforcement learning researcher would easily recognize that the GTD algorithm's objective (left) is exactly the same as the one on the right. The only difference is I wrote in the matrix-format and showed for the law-of-large-number estimates of A and b, while the GTD paper used the expectation format. Both are equivalent in a straightforward sense. I also proposed the gradient descent approach for this objective function. It is $O(n^2)$ though, to be clear.

For decades, a convergent temporal difference learning algorithm is missing (with a peak of arguments around 1995). Baird also proposed the residual gradient algorithm which performs some sort of gradient descent for Bellman error. However, it is not guaranteed to the TD solution. The objective, first proposed by the preconditioning paper, instead, guarantees the convergence of the gradient descent procedure (of minimizing it) to the TD solution.

The key idea of the objective is to introduce symmetry, for which the TD algorithm's A is not symmetric. As can be seen, unfortunately, the GTD(0) algorithm did not credit this objective to the preconditioning paper. Instead, the GTD paper claimed novelty of this objective function as shown in Figure 2. The TDC paper named this objective function (shown in Figure 3). According to the convention of the AI research community, naming something (an objective function, an algorithm or a model) takes explicit authorship of the idea discovered. Future papers usually cite the idea using the name they invented.

As written by the GTD authors, using this objective function is a "distinctive feature" of their algorithm, and "This last equation (equation 6) is key to our analysis". This is true because the key of the analysis is the stability of the O.D.E, which is shown by equation (4) in the preconditioning paper.

GTD is a stochastic approximation or optimization algorithm. For any optimization algorithm (e.g., LASSO), the objective function is the key. I feel it's boring for me taking my time and Rich's time as well on these obvious things. So I stop here and not continue writing about the TDC paper. I just paste the empirical results in the preconditioning paper in Figure 4, and TDC paper explained itself in Figure 5. In one sentence, TDC exploits both the ideas of symmetry and preconditioning.

Given the above evidence, it's safe to conclude that the GTD and TDC papers were both inspired by the preconditioning paper. These two paper's major contribution is to apply the objective proposed therein to off-policy learning and has a genius way (two-time scale approximation is used in a smart way here) of reducing the complexity to O(n) with a rigorous convergence proof. However, the omission of the credit to the preconditioning paper obscures the source of the idea and technical developments in this line of research, which have confused past readers and will continue to mislead future readers if I still remain silent. I'm pretty sure this note is a surprise for some of the coauthors of GTD and TDC papers too.

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 2: The GTD paper claimed this objective function is "novel".

Finally, we close this discussion of objective functions by giving the function used to derive the original GTD algorithm. This objective function does not seem to have a ready geometric interpretation. Here we call it the *norm of the expected TD update*:

$$\text{NEU}(\theta) = \mathbb{E}[\delta\phi]^\top \mathbb{E}[\delta\phi]\,. \qquad (6)$$

Figure 3: The TDC paper named this objective function.

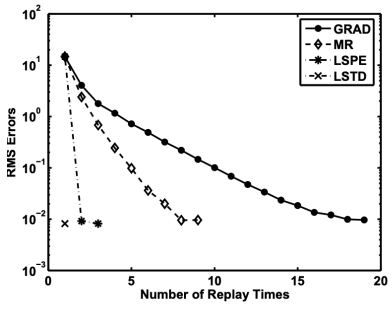


Figure 4. The role of preconditioner ($\lambda = 1$). Algorithms are stopped if RMS error is smaller than 0.01.

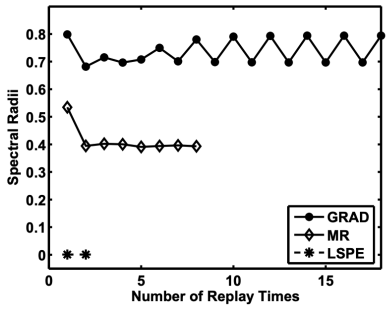
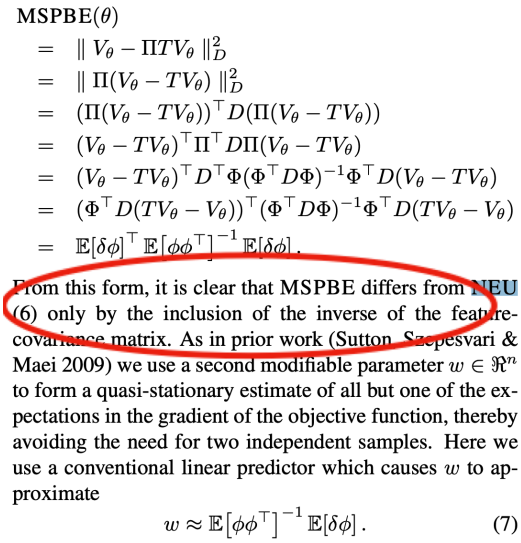


Figure 5. Spectral radius comparisons ($\lambda = 1$). LSTD's spectral radius is 0 permanently, thus not shown.

Figure 4: This figure in fact shows that GTD is slow (in the steady-state sense). MR is the faster version of iLSTD. iLSTD is the steady-state version of TD (TD cannot be faster than iLSTD). Check the abstract of the TDC paper on what was found.



$$
\begin{aligned}
&\text{MSPBE}(\theta) \\
&= \| V_\theta - \Pi T V_\theta \|_D^2 \\
&= \| \Pi(V_\theta - T V_\theta) \|_D^2 \\
&= (\Pi(V_\theta - T V_\theta))^\top D (\Pi(V_\theta - T V_\theta)) \\
&= (V_\theta - T V_\theta)^\top \Pi^\top D \Pi (V_\theta - T V_\theta) \\
&= (V_\theta - T V_\theta)^\top D^\top \Phi (\Phi^\top D \Phi)^{-1} \Phi^\top D (V_\theta - T V_\theta) \\
&= (\Phi^\top D (T V_\theta - V_\theta))^\top (\Phi^\top D \Phi)^{-1} \Phi^\top D (T V_\theta - V_\theta) \\
&= \mathbb{E}[\delta\phi]^\top \mathbb{E}\left[\phi\phi^\top\right]^{-1} \mathbb{E}[\delta\phi] .
\end{aligned}
$$

From this form, it is clear that MSPBE differs from NEU (6) only by the inclusion of the inverse of the feature-covariance matrix. As in prior work (Sutton, Szepesvari & Maei 2009) we use a second modifiable parameter $w \in \Re^n$ to form a quasi-stationary estimate of all but one of the expectations in the gradient of the objective function, thereby avoiding the need for two independent samples. Here we use a conventional linear predictor which causes w to approximate

$$w \approx \mathbb{E}\left[\phi\phi^\top\right]^{-1} \mathbb{E}[\delta\phi] . \tag{7}$$

Figure 5: The TDC paper explained how TDC was developed.

3 Some sort of conclusion. or no conclusion. Science needs to move on.

I never imagined I would write down this any time since 2008. I had chosen to take easy about it. I had chosen to focus on importance things and ignore this. I can live without a credit for GTD and TDC. Science gets better, anyway. No matter who had the ideas first. No matter the original thinker was credited or not. I suspect this happened a lot in the history of science, especially in the old days of mathematics, when communication and publishing was a problem. These days, perhaps cases like this are extremely rare. As a reviewer for the AI community for many years, I can say that what I experienced is generally healthy and positive, whether I was an author or a reviewer. In some cases, biased and unfair reviews can happen; but once raised, it is usually taken care of. In some case, it was not taken care of, but the rate of this is low. No system is perfect. The preconditioning paper was reviewed highly by two reviews (one reviewer said something like "this paper contains brilliant ideas"). It was NIPS 2007, when they also gave a confidence score for their reviews. These two folks gave confidence like 9 and 7 out of 10. The third reviewer gave a score of 5 with a confidence of 2. Yet, the preconditioning paper was rejected. The next time I submitted, ICML 2008, all the three reviewers gave an acceptance. The system was much better from then I think. I heard from someone that that year, they've seen papers have seven reviewers, in case there is a large deviation in the opinions, they just keep adding reviewers until agreement is reached. Honestly, I finally brought this up not because of credit or authorship thing, even though it was important for me. Only recently, I realized how hard it was for me for my studentship.

Something struck me in January 8th, 2020. On that date, Ukraine International Airlines Flight 752 was shot down minutes after taking off from Tehran, Iran. On that flight, there carried Arash my mentor student, born in 1993, and his wife. On that date hit by the news (or Paniz came to office and told me? I forgot), I called members in the office (a few intern students and a few Huawei employees) together and held a few minutes of silence for Arash and the other victims. I remembered I sobbed, which I couldn't hold.

I didn't know why. I cried again on the memorial for the victims held by UofA in South Campus. I brought all my kids and told them to put flowers on Arash and his wife's table. My daughter was talking to me on something. My son Alex shouted, "Mim! Don't you see it? Daddy is crying and he is sad!" Arash and his wife, and the other victims, their family and their smiles just couldn't let it go off my mind after a long, long time.

Doing research is a privilege. Writing about a research topic is a privilege too. A student looked up to her/his supervisor when she/he came to learn, not only the knowledge of science but also the spirit of science. I used to blame myself badly for not having worked hard enough and not having

done good enough during my Ph.D studies, especially the first three years with Rich. I used to blame myself for choosing to do PageRank, which I did not succeed, and giving up my loved reinforcement learning since 2010 and for a few years. When I finally realized PageRank was not going to work for my Ph.D thesis, I came back to and asked Rich whether I could still defend on reinforcement learning. I remembered he said something like "choices have consequences". I'm sorry to mention to this but Rich said was right. However, it just sound hard to me at the time. I knew I perhaps made an irrational choice giving up reinforcement learning on my own and Rich indeed strongly suggested me not to at the time. Perhaps Rich said this just because I didn't listen to him on his advice. I really appreciate that. Later Csaba told me Rich didn't want to be part of my defence committee. I fully understood and understand that. I didn't make a smart choice on my own and I can't expect everyone responds the same and to what one expects. I was still lucky enough that Csaba and Dale supported me and I didn't walk away without a Ph.D. This enormously helped my career and my family. I moved on.

Only recently, when writing a work on step-size planning and reviewing linear Dyna, I recalled a senior researcher gently reminded me in 2017 he was upset that linear Dyna didn't cite his work (he never explicitly said this, but I believe he was implying linear Dyna, "someone derived ideas from my paper but without citing it"). He has way many papers and books. I looked around his office. Two bookshelves of books, "these are all the books that I wrote." He said to me gently, looking at me with his bright and intelligent eyes, without even a slight complacency or arrogance on his face. He is very famous. He is very senior. Yet he cared. Being recognized by the community for one's contribution is all what we cherish regardless of senior or young researchers. He found a book, turned to the main text and reference, and pointed to me, "here I cite your paper". I never have a feeling like that. It's like the feeling is much better than even getting accepted by ICML for the first time. We are not PDF machines, we are humans, right? So this old memory revisited me, and urged me to tell Rich and his coauthors about this. In the cases of GTD and TDC/GTD2, things are even more clear. Yet, I chose to ignore and not care. but these thoughts came back, you know, like "why should I not care?" I'd chosen to not care, yes, for many years. I'd chosen to give up reinforcement learning, yes, for a few years. TDC is not the last paper Rich wrote on this topic. There are GQ algorithms. There are the second version of the RL book. None of these mentioned my preconditioning paper on the GTD section unfortunately.

Only recently, I figured out my studentship was not even quite right in the very beginning. I wasn't even in the lab yet when the GTD paper was written. Rich is someone who I extremely respect and who I have been and I still am so much influenced by. Check my step-size planning and you will be surprised by the number of times that I cited him and his coauthors. (Tell me if you found a citation to him there does not concern an idea or a technical matter). This is also another reason I had chosen to remain in silence about this matter. Rich is famous. He didn't need two or more papers to elevate his fame. However, the way he wrote, especially on the originality of the ideas, perhaps can be improved, with a vigorous review of others' ideas especially similar ones, not just obsessed in his own ideas and deliver them in even slightly obscure way.

The GTD and TDC/GTD2 papers are still great and ground breaking papers. I cited them in my later papers. Rich is a key figure in reinforcement learning. I believe he did this unintentionally and he probably thought it's just Okay because my paper is focused on on-policy learning and his are instead on off-policy learning. I totally agree and get this. This case is just a bit special: mathematically, the objective function, transfers from on-policy learning to off-policy learning without any difficulty, in fact, the derivation is exactly the same. Did they find something different in the convergence rate in the off-policy learning case from on-policy, like the rates shown in Figure 4? Do you notice the title of TDC, "Fast Gradient-Descent Methods ...", why TDC is faster than GTD and did the TDC paper explain why? And why were they so confident in using the title "Fast Gradient-Descent Methods" without explaining the technical reason? These are my technical questions if I was a reviewer.

I only wish Rich had a serious thought about the implication for a young student who came for him by reading his TD prediction paper and the RL literature for three years, and the series of possible consequences for the student, some of which unfortunately indeed happened to me. Research are not pdfs of equations, figures and tables. Behind every paper, there is thinking, creativity, experimentation, hopelessly giving up, recharging and starting over; all like that. For the objective function, gradient descent idea for TD and preconditioning, it was a hot summer in Shenzhen, in a weekend on the bed running my pen on stacks of paper sheets in a small dormitory where we rent at the time. The preconditioning paper was submitted four times since 2006. One could say it's not hard to derive

this objective function. Yes. But it takes you ten minutes to get the idea once someone showed you, and it took decades of the reinforcement learning community, e.g., Baird's residual gradient in 1995 and before that, Tsitsiklis, Van Roy, Gordon, Moore, Boyan, Sutton, Precup, Bertsekas (remember his counter example to TD?), Tadic, Dayan, and perhaps more researchers that I don't know, all did lots of work on the convergence and divergence of TD. This is why GTD is ground breaking. Is it not ground breaking if had one cited the source of the idea in an important paper? I think a paper is ground breaking or not does not depend on the source of the idea is proposed by the authors or not. The Adam paper in deep learning used the momentum technique, which was started by Polyak at least no later than 1980s. In this case, $O(n^2)$ methods there are many: LSTD, LSPE, and my gradient TD method. The genius way that GTD stands for never existed in literature and it had bothered the RL community for such a long time. To Rich: I think it's not a shame to cite your student for the source of idea even though he was just a master student when writing the paper. I had hard times of understanding your motivation. I can only assure myself this was an unintentional mistake.

I particularly think this note is a gentle reminder to famous people. They all have worked their hard ways with good research most of the cases a number of seminal papers. Because of their track records, key people are exponentially easier to get traction after their flat and linear curves. Unintentional obscurity in research could cause very detrimental effects for junior researchers who are young and creative. The fact that they are young means they don't know how to deal with it rationally, they might think that's the only paper she or he will ever have (which is not), nor did they have the knowledge or experience, and it could also mean unnecessary mental pressure for them given that they have very little recognition by the research community. A good system of science is perhaps to cite young students and researchers with more attention, and let them focus on the important questions and move science on.

If anything that is worth your time of reading until here —

One should always watch for examples like discussed in this note slip into your research.