Policy and Value Based Learning

Statistical & Convergence Properties

Questions we’re trying to answer

- What are the statistical properties of both approaches?
- What’s the behavior of the models in large action spaces?
- Do we need complex models to get good results?
- How easy is the optimization of the objectives?
- Can we have the benefits of both worlds?

Setup :

We have an online system interacting with users :

- Users are represented with contexts x
- The system is represented by a stochastic policy π_0, that given x, does action a (recommend a product) and get a feedback $r(a, x)$ (sale, click..)

→ We are interested in finding the policy that maximizes the expected reward :

$$\pi^* = \; argmax_{\pi} \, \mathbb{E}_{\pi,x,r}[r(x,a)] = argmax_{\pi} \, R(\pi)$$

The Assumption :

The online system is still exploring, so for any user, all the actions can be done.

We say that the policy of our system has ***full support over the actions***

$$\tau = \min_{\{x,a\}} \pi_0(a|x) > 0$$

Q1 – Statistical Properties?

- Value based approach

$$R(\pi) \approx \hat{R}_{model}(\pi) = \frac{1}{N}\sum_i \hat{r}_\theta(a_i, x_i)\pi(a_i|x_i)$$

  - The parametrization is on the reward → $\pi(a|x) = 1[a = argmax_a \hat{r}_\theta(a_i, x_i)]$
  - Natural Bias- variance trade-off coming from the modelling part.

  For example, the estimator is unbiased if the reward model is well-specified.

- Policy based approach with IPS : inverse propensity scoring

$$R(\pi) \approx \hat{R}_{IPS}(\pi_\theta) = \frac{1}{N}\sum_i r(a_i, x_i)\frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)}$$

  - The parametrization is on the policy → Look for $\mathrm{argmax}_\theta \hat{R}(\pi_\theta)$
  - Unbiased if π_0 has full support. The variance of the estimator depends on the distance between π_θ and π_0.

  → Unbiased but can suffer from huge variance, other estimators provide a better bias-variance tradeoff.

  See for example cIPS[1], SNIPS [2], DR[3], SWITCH[4]…

Some Preliminary Results

- Under some assumptions**, Statistical learning theory** provides insight on how close can we get to the optimal policy with value based or policy-based models in terms of :

  - The complexity of the model $\mathcal{C}(M)$: the more complex the model, the bigger the value of $\mathcal{C}(M)$
  - The number of datapoints we gathered.
  - The stochasticity of our logging policy/previous policy of the recommender system.

- The following results come from both [5][6] :

  - ***Value based*** : If the model is ***well-specified***, we reach the optimal policy at a rate of $\mathcal{O}(\sqrt{\frac{\mathcal{C}(M)}{N\tau}})$
  - ***Policy based*** : It is guaranteed to return nearly the optimal policy at a rate of $\mathcal{O}(\frac{1}{\tau}\sqrt{\frac{\mathcal{C}(M)}{N}})$

Q2 – When the action space is large?

- ***Value based*** : If the model is ***well-specified***, we reach the optimal policy at a rate of $\mathcal{O}(\sqrt{\frac{\mathcal{C}(M)}{N\tau}})$
- ***Policy based*** : It is guaranteed to return nearly the optimal policy at a rate of $\mathcal{O}(\frac{1}{\tau}\sqrt{\frac{\mathcal{C}(M)}{N}})$

As you might remember, $\tau = \min_{\{x,a\}} \pi_0(a|x) \leq \frac{1}{\mathcal{A}}$ with equality when π_0 is uniform.

→ τ gets smaller when the action space is big.

Best case scenario : $\tau = \frac{1}{\mathcal{A}}$

$\mathcal{O}(\mathcal{A}\sqrt{\frac{\mathcal{C}(M)}{N}})$ for ***Policy*** **-** $\mathcal{O}(\sqrt{\mathcal{A}\frac{\mathcal{C}(M)}{N}})$ for ***Value***

→ Value based approach behave better in large action spaces if the model is ***well-specified.***

Q3 – The need for complex models?

- **Value based Learning** results were obtained with the assumption that the model is ***well-specified***.
- In the real world, this assumption can hold to a certain degree if :
  - we use powerful models. (Deep NNs, Gaussian Processes..)
  - we have domain-specific expertise (Bayesian Hierarchical models..)

  → In ***the general case***, we need complex models for Value based approches, the rate $\mathcal{O}(\sqrt{\frac{\mathcal{C}(M)}{N\tau}})$ tend to increase.
- **Policy based learning** doesn't need this assumption to achieve its rate of $\mathcal{O}(\frac{1}{\tau}\sqrt{\frac{\mathcal{C}(M)}{N}})$

Empirically :

- Policy based learning tend to outperform Value based learning with simple linear functions. See [3] for example.
- Value based approaches outperform policy when we have a good model of the reward. See BLOB [7] for the case of personnalized advertising.

Q4 – Objectives easy to optimize?

- Value based learning can be cast into ***a regression problem***.
  - Multitude of convex objectives for the simple linear case (Generalized Linear Models...)
  - Leverage the success of Deep Neural Networks/Variational Inference for complex models.

- Even in the simple linear case, Policy Based learning has a ***highly non convex objective*** [9]

Theorem 1 *Even for a single context x, a deterministic reward vector $\boldsymbol{r}$, and a linear model $\boldsymbol{q}(x) = W\boldsymbol{\phi}(x)$, the function $\boldsymbol{r} \cdot \boldsymbol{f}(\boldsymbol{q}(x))$ can have a number of local maxima in W that is exponential in the number of actions K and the number of features in $\boldsymbol{\phi}$.*

Practical Guidelines?

- If you think you understand well the phenomenon, and you deal with very large action spaces, one can go for value based methods as it provides good results.

- If the reward is hard to model, and you deal with small action spaces, **Policy Based learning** is more suitable.

- The case where the reward is hard to model and we deal with huge actions spaces is still an active area of research, and one can :

  - Use better estimators/learning objectives for Policy Based learning. See for example [1][2][11][12].
  - ***Combine*** reward modelling and Policy based learning in the hope of getting better results.

NB : the size of the action space is measured relatively to the number of observations one has/complexity of the model.

Q5 – Having the benefits of both worlds?

- Use a reward model with Policy learning to achieve good ***variance reduction***, as an example:
  - [3, ***Doubly Robust Policy Evaluation & Learning***] : Use a reward model as a **control variate** to reduce variance.
  - [4, ***Optimal and Adaptive Off-policy Evaluation in Contextual Bandits***] : Switches between IPS and a reward model-based estimator depending on how far the current policy is from the logging policy to achieve better bias-variance trade-off.
  - [10, ***CAB: Continuous Adaptive Blending for Policy Evaluation and Learning***] : Adaptively Blending different estimators (policy-based and reward-model based) to achieve optimal mean squared error.

- Training jointly a reward model with a policy model to ***convexify*** the learning objectives :
  - [9, ***Surrogate Objectives for Batch Policy Optimization in One-step Decision Making***] : Defines a ***convex*** (in the linear case) calibrated surrogate loss to the policy based objective which learns jointly a reward model to converge to a better optima.
  - [8, ***Joint Policy-Value Learning for Recommendation***] : Defines a ***convex*** (in the linear case) upper bound on the policy objective with the help of a negative log-likelihood (model learning) to have a better learning objective.

It’s time to dive into the practical session!

References

- [1] Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. **Counterfactual reasoning and learning systems: The example of computational advertising**. The Journal of Machine Learning Research, 14(1):3207–3260, 2013.
- [2] Swaminathan, A. and Joachims, T. **The self-normalized estimator for counterfactual learning**. In advances in neural information processing systems, pp. 3231–3239, 2015b.
- [3] Dudík, M., Langford, J., and Li, L. **Doubly robust policy evaluation and learning**. arXiv preprint arXiv:1103.4601, 2011.
- [4] Wang, Y.-X., Agarwal, A., and Dudik, M. **Optimal and adaptive off-policy evaluation in contextual bandits**. In International Conference on Machine Learning (ICML), 2017.
- [5] Strehl, A., Langford, J., Li, L., and Kakade, S. M. **Learning from logged implicit exploration data**. In Advances in Neural Information Processing Systems, pp. 2217–2225, 2010.
- [6] Chen, J. and Jiang, N. **Information-theoretic considerations in batch reinforcement learning**. In Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019.
- [7] O. Sakhi, S. Bonner, D. Rohde, and F. Vasile. 2020. **BLOB : A Probabilistic Model for Recommendation that Combines Organic and Bandit Signals**. In Proc. of the 26th ACM Conference on Knowledge Discovery & Data Mining (KDD ’20).
- [8] O. Jeunen, D. Rohde, F. Vasile, and M. Bompaire. 2020. **Joint Policy-Value Learning for Recommendation**. In Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD ’20).
- [9] Chen, M., Gummadi, R., Harris, C., and Schuurmans, D. **Surrogate objectives for batch policy optimization in one-step decision making**. In Advances in Neural Information Processing Systems, pp. 8825–8835, 2019
- [10] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims**. CAB: continuous adaptive blending for policy evaluation and learning**. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 6005–6014. PMLR, 2019.
- [11] Swaminathan, A. and Joachims, T. **Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization**. Journal of Machine Learning Research, 16:1731–1755, 2015a.
- [12] Louis Faury, Ugo Tanielan, Flavian Vasile, Elena Smirnova, and Elvis Dohmatob. 2020. **Distributionally Robust Counterfactual Risk Minimization**. In Thirty-Fourth AAAI Conference on Artificial Intelligence.