

W20: Learning from Logged Implicit Exploration Data

Alexander L. Strehl, John Langford,
Lihong Li, Sham Kakade

Offline Policy Learning Problem

extra dark chocolate (About this page) - 0.20 sec.

SPONSOR RESULTS

[Extra Dark Chocolate](#)

Shop 80,000+ products with one cart. Your online Gourmet Food source.

Amazon.com/Gourmet

[Fresh Dark Chocolate](#)

Fresh gourmet **dark chocolate** sure to astound. Truffles, caramels,...

www.lakechamplainchocolates.com

[Chocolate by Marky's - Dark Chocolate](#)

Leonidas Belgian **chocolate** gourmet gifts mail order online.

www.markys.com

[A Lindt Extra Dark Chocolate](#)

Buy a Lindt **Extra Dark Chocolate** at SHOP.COM.

www.SHOP.com

First month's policy

Second month's policy

Can we find a better policy given click logs from each policy?

extra dark chocolate (About this page) - 0.20 sec.

SPONSOR RESULTS

[A Lindt Extra Dark Chocolate](#)

Buy a Lindt **Extra Dark Chocolate** at SHOP.COM.

www.SHOP.com

[Fresh Dark Chocolate](#)

Fresh gourmet **dark chocolate** sure to astound. Truffles, caramels,...

www.lakechamplainchocolates.com

[Chocolate by Marky's - Dark Chocolate](#)

Leonidas Belgian **chocolate** gourmet gifts mail order online.

www.markys.com

[Extra Dark Chocolate](#)

Shop 80,000+ products with one cart. Your online Gourmet Food source.

Amazon.com/Gourmet

Formalization

- Given **logged data** from a **production system** π , evaluate the performance of a new and **different** policy.
- Our training examples are of the form (Input, Action, Reward), where actions were chosen by π , and rewards for unknown actions are not revealed.
- This is different (and harder) than **supervised learning** or **contextual k-armed bandits**.

Solution, Results, Experiments

- Our **offline policy estimation** takes two steps:
 - Model the production system by conditional probability estimation.
 - Using this model, evaluate a new policy via importance sampling to correct for bias of production system.
- We prove that it converges to a biased estimate and provide sample complexity bounds for using it for policy optimization.
- We present experiment results on it's application to **Yahoo!'s Front Page** and **Online Advertising** products.