Per-Decision Option Discounting

Anna Harutyunyan, Peter Vrancx,
Philippe Hamel, Ann Nowe,
Doina Precup

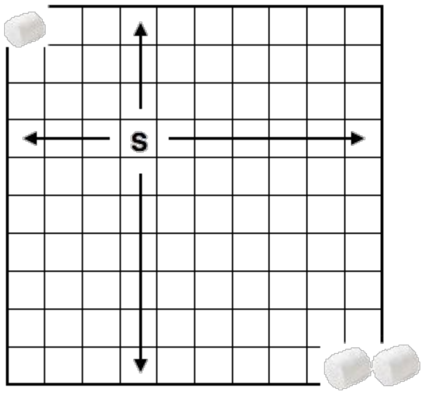
VRIJE UNIVERSITEIT BRUSSEL
DeepMind
PROWLER.io the decision company

Motivation: Agents that reason over long temporal horizons

Motivation: Agents that reason over long temporal horizons

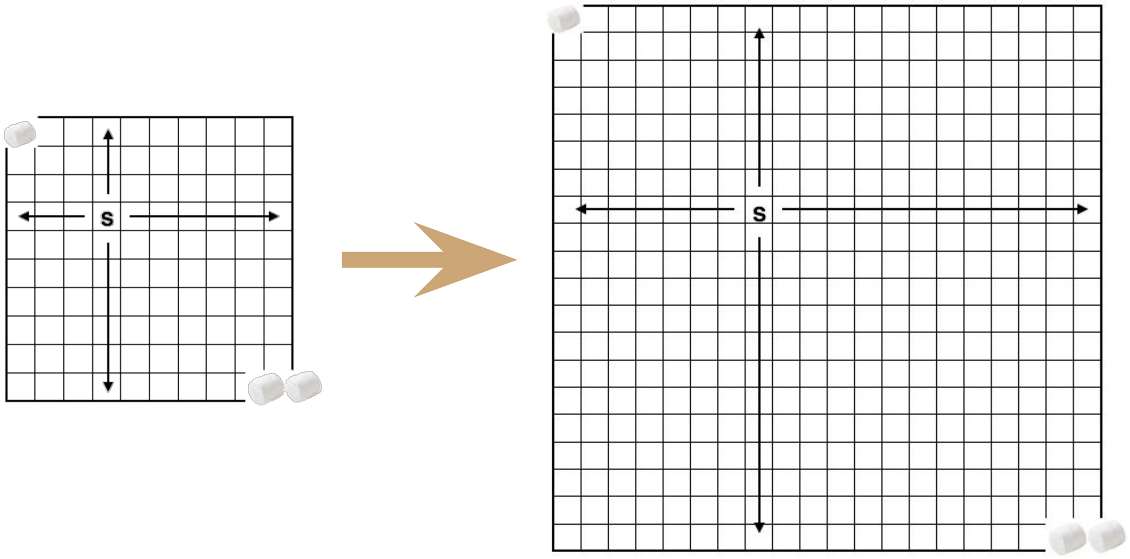
Horizon depends on discount γ

Motivation: Agents that reason over long temporal horizons



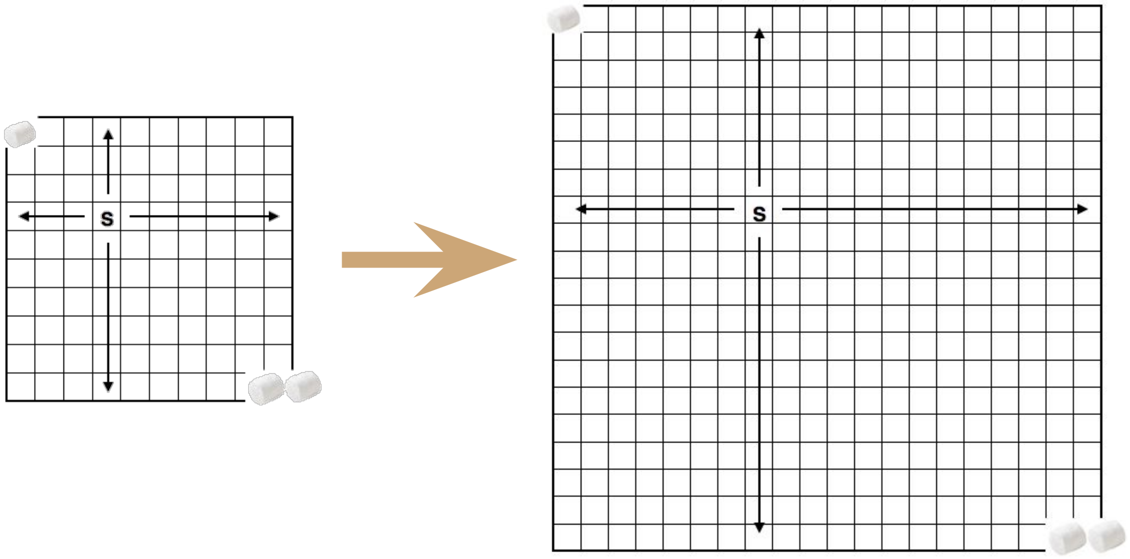
Horizon depends on discount γ

Motivation: Agents that reason over long temporal horizons



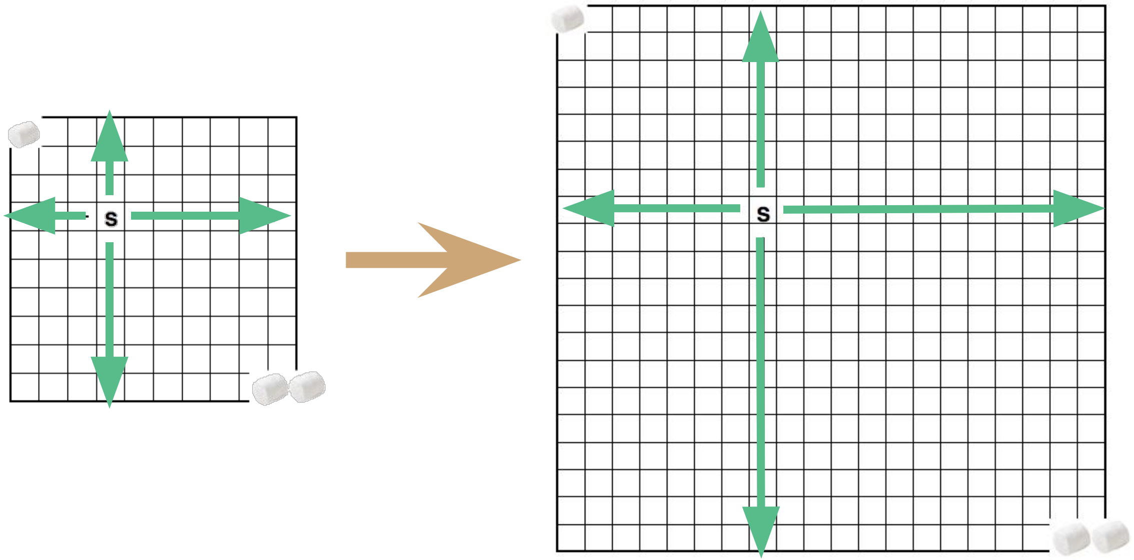
Horizon depends on discount γ
Larger grid requires a larger γ

Motivation: Agents that reason over long temporal horizons



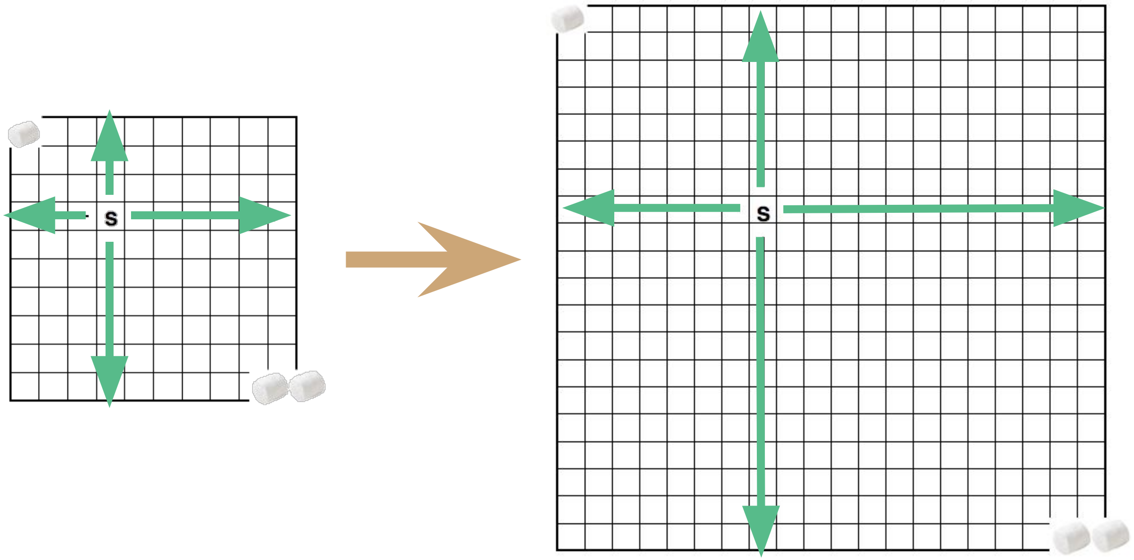
Horizon depends on discount γ
Larger grid requires a larger γ
Large γ-s are inefficient in practice :(

Motivation: Agents that reason over long temporal horizons



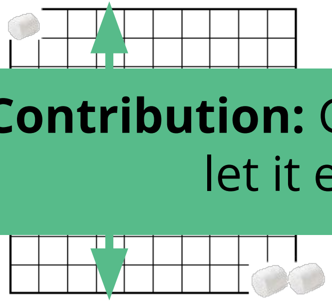
Horizon depends on discount γ
Larger grid requires a larger γ
Temporal abstraction?

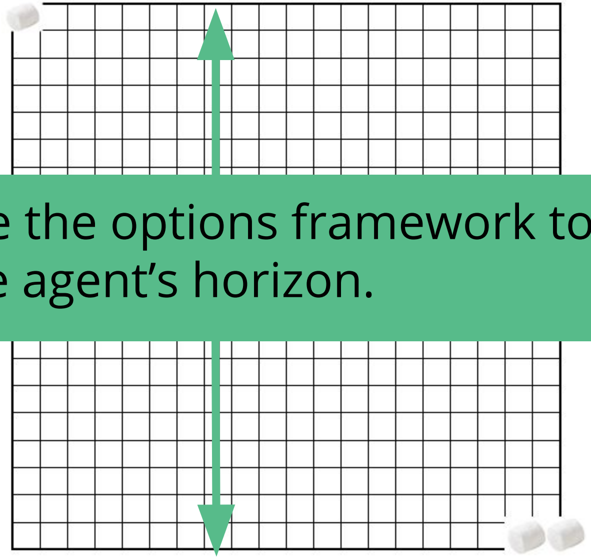
Motivation: Agents that reason over long temporal horizons



Horizon depends on discount γ
Larger grid requires a larger γ
Temporal abstraction? Options still tied to γ!

Motivation: Agents that reason over long temporal horizons





Contribution: Generalize the options framework to let it extend the agent's horizon.

Horizon depends on discount γ

Larger grid requires a larger γ

Temporal abstraction? Options still tied to γ!

The Options Framework

Reward model:

$$R^o(s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s|o}\Big[\sum_{t=0}^{D-1} \gamma^t R_{t+1}\Big]$$

Transition model:

$$P^o(s'|s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s\rightarrow s'|o}\Big[\gamma^D \mathbb{I}_{S_D=s'}\Big]$$

The Options Framework

Reward model:

$$R^o(s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s|o}\left[\sum_{t=0}^{D-1} \gamma^t R_{t+1}\right]$$

Transition model:

$$P^o(s'|s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s\to s'|o}\left[\gamma^D \cdot S_{D=s'}\right]$$

Options with Time Dilation

Reward model:

$$R^o(s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s|o}\Big[\sum_{t=0}^{D-1} \gamma^t R_{t+1}\Big]$$

$$R^o(s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s|o}\Big[\sum_{t=0}^{D-1} \gamma_r^t R_{t+1}\Big]$$

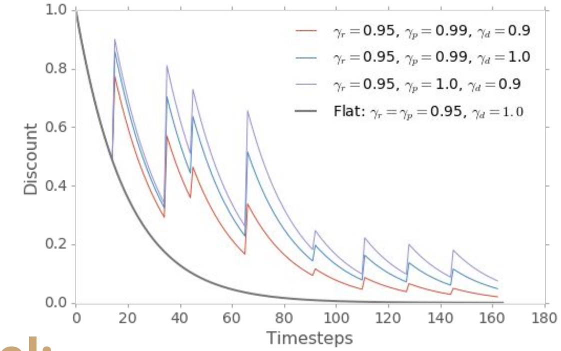
Transition model:

$$P^o(s'|s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s\to s'|o}\big[\gamma^D \mathbb{1}_{S_D=s'}\big]$$

(2) per-decision (1) decouple

$$P^o(s'|s) \stackrel{\text{def}}{=} \gamma_d \mathbb{E}_{D:s\to s'|o}\big[\gamma_p^D \mathbb{1}_{S_D=s'}\big]$$

Options with Time Dilation



Reward model:

$$R^o(s) \overset{\text{def}}{=} \mathbb{E}_{D:s|o}\left[\sum_{t=0}^{D-1} \gamma^t R_{t+1}\right]$$

↓

$$R^o(s) \overset{\text{def}}{=} \mathbb{E}_{D:s|o}\left[\sum_{t=0}^{D-1} \gamma_r^t R_{t+1}\right]$$

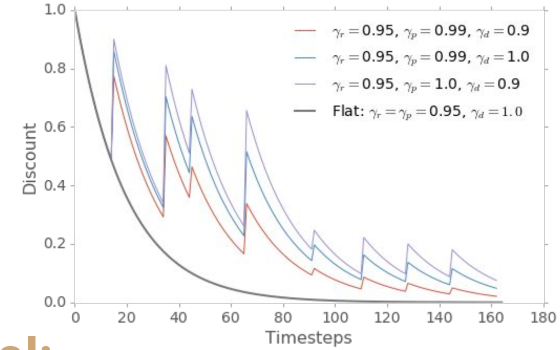
Transition model:

$$P^o(s'|s) \overset{\text{def}}{=} \mathbb{E}_{D:s\to s'|o}\left[\gamma^D \mathbb{1}_{S_D=s'}\right]$$

↓ (2) per-decision (1) decouple

$$P^o(s'|s) \overset{\text{def}}{=} \gamma_d \mathbb{E}_{D:s\to s'|o}\left[\gamma_p^D \mathbb{1}_{S_D=s'}\right]$$

Options with Time Dilation



Reward model:

$$R^o(s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s|o}\Big[\sum_{t=0}^{D-1}\gamma^t R_{t+1}\Big]$$

$$R^o(s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s|o}\Big[\sum_{t=0}^{D-1}\gamma_r^t R_{t+1}\Big]$$

Transition model:

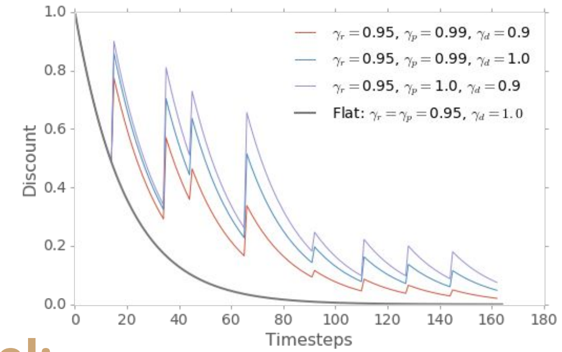
$$P^o(s'|s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s\rightarrow s'|o}\Big[\gamma^D \mathbb{1}_{S_D=s'}\Big]$$

(2) per-decision (1) decouple

$$P^o(s'|s) \stackrel{\text{def}}{=} \gamma_d \mathbb{E}_{D:s\rightarrow s'|o}\Big[\gamma_p^D \mathbb{1}_{S_D=s'}\Big]$$

γ_p controls how much we care about option duration (pseudo-primitive when γ_p=1)

Options with Time Dilation



Reward model:

$$R^o(s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s|o}\left[\sum_{t=0}^{D-1} \gamma^t R_{t+1}\right]$$

↓

$$R^o(s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s|o}\left[\sum_{t=0}^{D-1} \gamma_r^t R_{t+1}\right]$$

Transition model:

$$P^o(s'|s) \stackrel{\text{def}}{=} \mathbb{E}_{D:s\rightarrow s'|o}\left[\gamma^D 1_{S_D=s'}\right]$$

↓

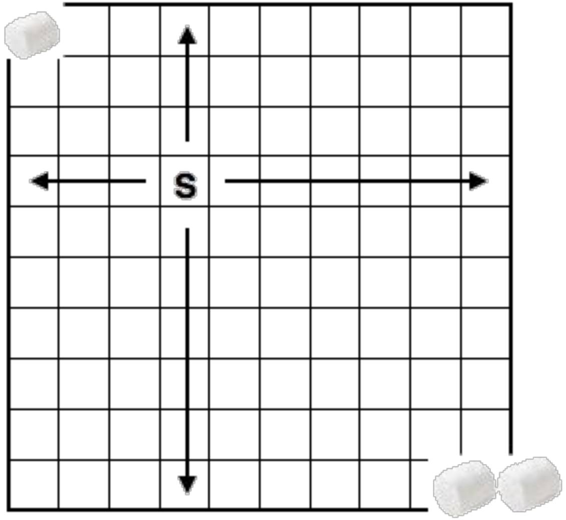
(2) per-decision (1) decouple

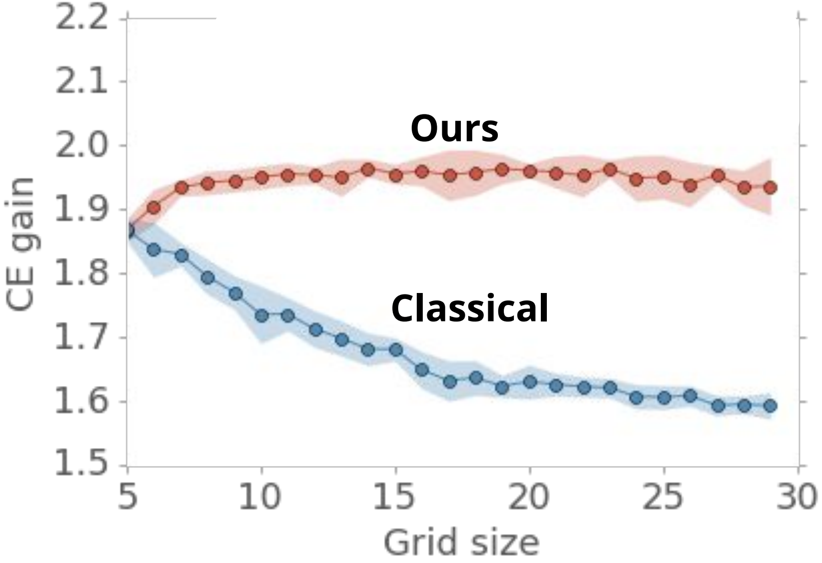
$$P^o(s'|s) \stackrel{\text{def}}{=} \gamma_d \mathbb{E}_{D:s\rightarrow s'|o}\left[\gamma_p^D 1_{S_D=s'}\right]$$

γ_p controls how much we care about option duration (pseudo-primitive when γ_p=1)

Key intuition: Insulate option time from global time

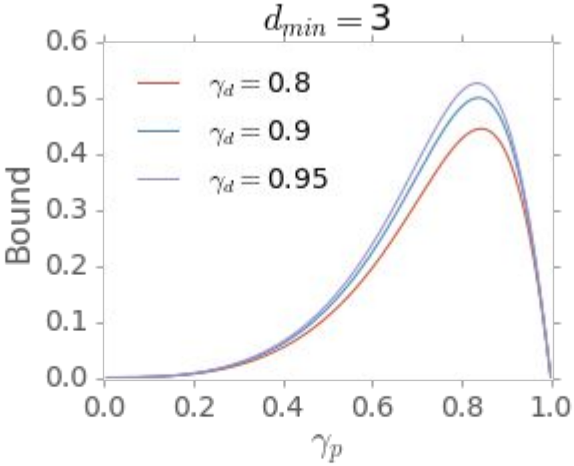
Primitive Timestep Invariance



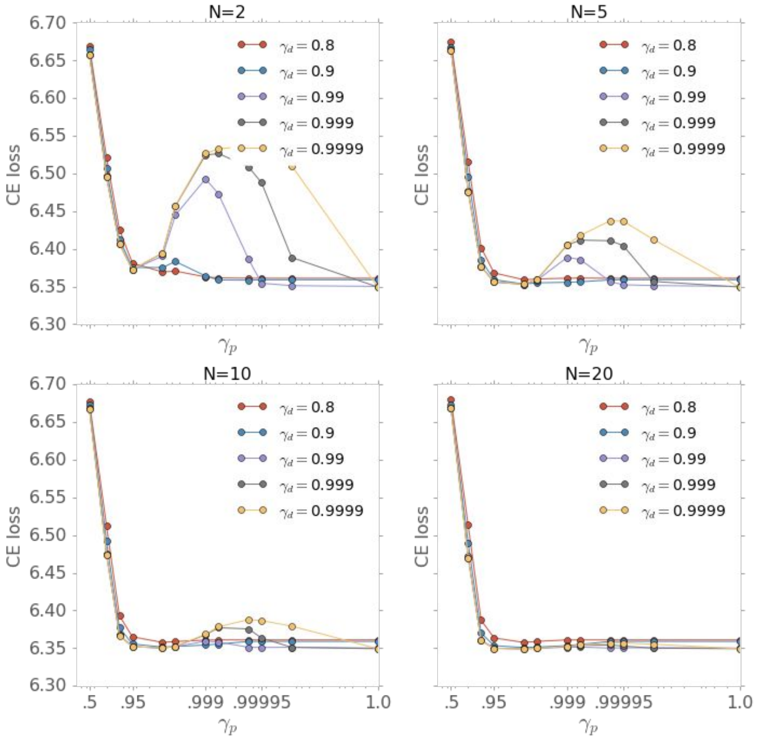


Bias-Variance Tradeoff

Analytical variance bound

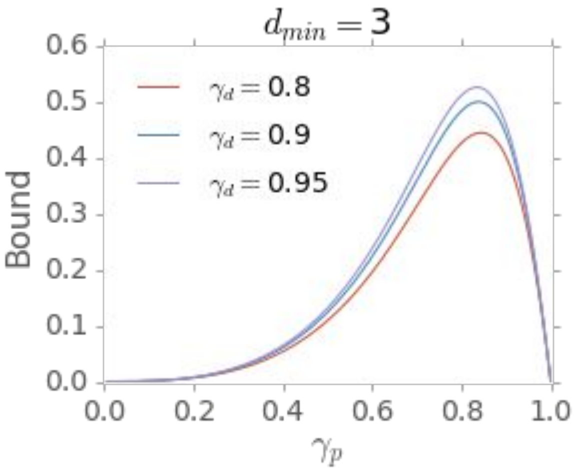


Empirical error (Four Rooms)



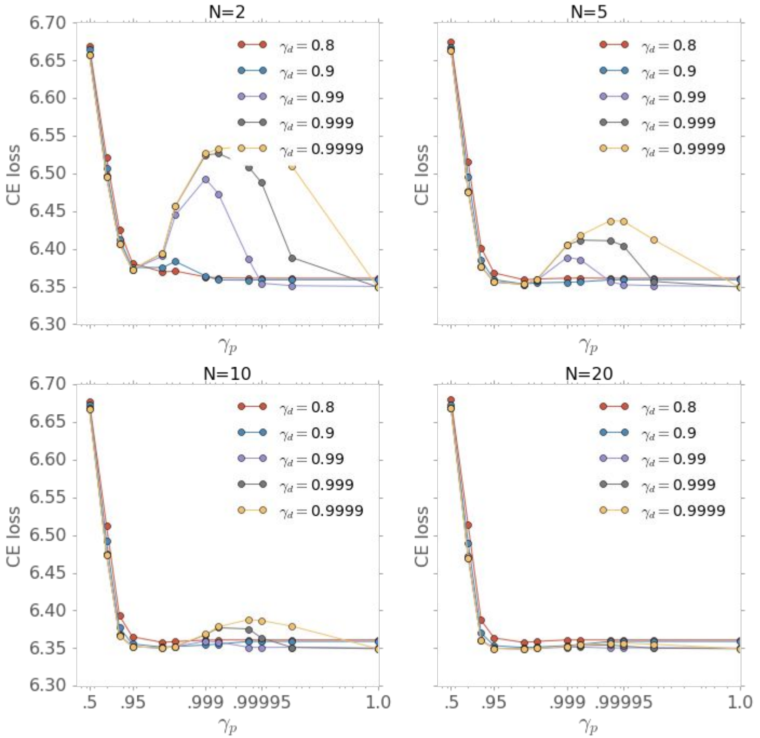
Bias-Variance Tradeoff

Analytical variance bound



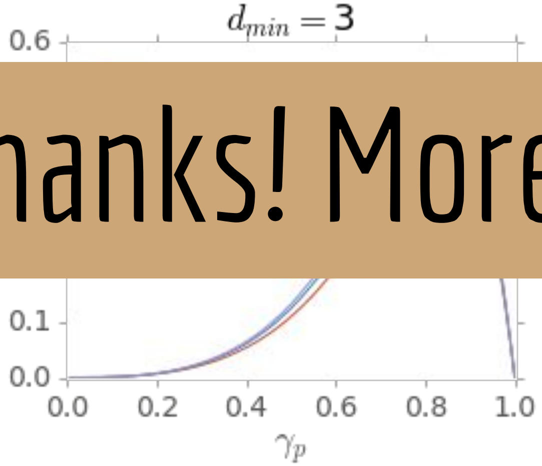
Larger *γ_p can induce* ***less*** *variance!*

Empirical error (Four Rooms)



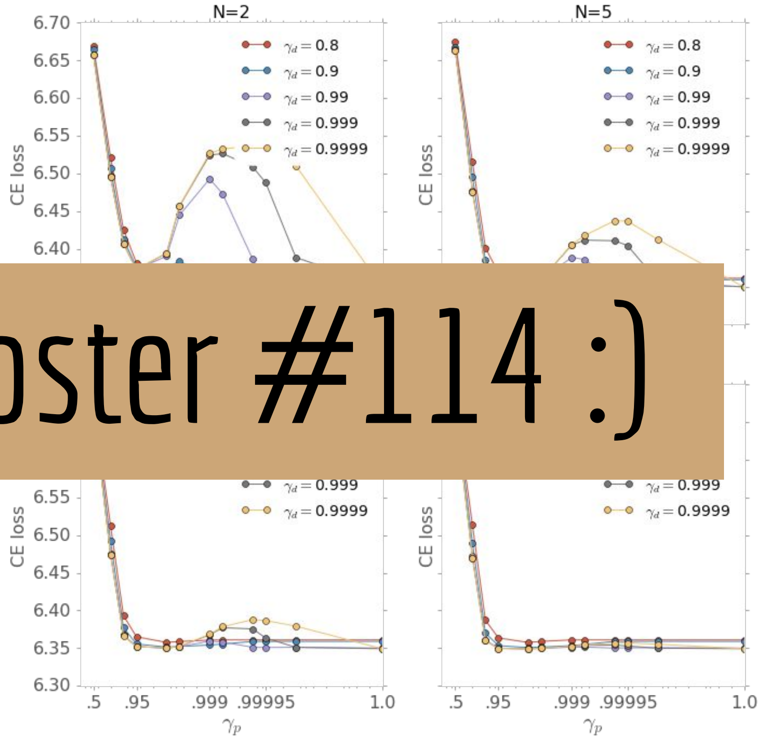
Bias-Variance Tradeoff

Analytical variance bound



Larger *γ_p can induce* ***less*** *variance!*

Empirical error (Four Rooms)



Thanks! More at poster #114 :)