Variance Driven Exploration

A Provable and Efficient Methodology for Pure Exploration in Highly Stochastic Environments

Khang Luong Nam Nguyen Hoang Ta Hung The Tran Tuan Dam

ICML 2026

Core message: sample where uncertainty most affects the final decision.

Pure exploration: the final decision matters

Setting

We collect noisy samples, then output a final object:

- best arm in bandits,
- best root action in tree search,
- near-optimal policy in RL.

Problem with local optimism

In highly stochastic or heteroscedastic environments, local confidence bonuses can chase noise or over-refine the wrong comparison.

Table: A stylized heteroscedastic BAI snapshot. Arm 1 is optimal in expectation but has larger variance. After a warm start, empirical ranking can favor the stable suboptimal arms, while decision-relevant uncertainty remains concentrated in Arm 1.

Arm	Distribution	μ_i	$\hat{\mu}_i$
1	$\mathrm{Bernoulli}(0.55)$	0.55	0.50
2	$\{0.48, 0.58\}$ equiprob.	0.53	0.53
3	$\{0.47, 0.57\}$ equiprob.	0.52	0.52

VarDE principle: minimize decision-level variance

Smooth final decision

$$Y = f(\hat{\mu}_1, \ldots, \hat{\mu}_n)$$

- $\hat{\mu}_i$: local empirical estimate
- Y: smooth surrogate of final decision

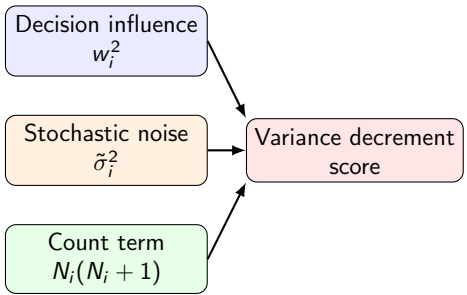
First-order influence

$$w_i(\mu) = \frac{\partial f}{\partial \hat{\mu}_i}(\mu)$$

$$\mathrm{Var}(Y_t) \approx \sum_i w_i(\mu)^2 \frac{\sigma_i^2}{N_i(t)}$$

Greedy sampling rule

$$i_t \in \arg\max_i \frac{w_i(\hat{\mu})^2 \, \tilde{\sigma}_i^2}{N_i(N_i+1)}$$



One principle, three algorithms

Problem	Smooth decision variable
BAI	$Y_\tau = \tau \log \sum_i \exp(\hat{\mu}_i/\tau)$
MCTS	$Y_\tau(s) = \tau \log \sum_{a \in A(s)} \exp(\hat{Q}(s,a)/\tau)$
BPI / Q-learning	$Y_\tau(s) = \tau \log \sum_{a \in A(s)} \exp(Q(s,a)/\tau), \quad y = r + \gamma \max_b Q(s',b)$

Guarantees: variance decay and correctness

Generic VarDE

Under regularity and boundedness assumptions:

$$\mathrm{Var}(Y_t) = O(t^{-1}).$$

The leading constant is minimized by

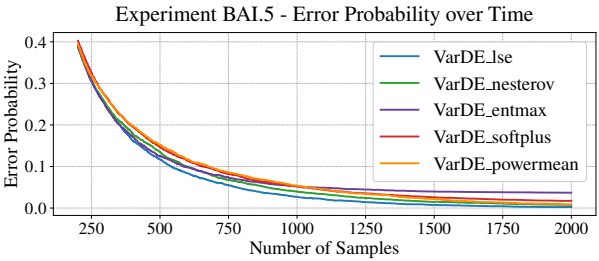
$$p_i^\star \propto |w_i(\mu)|\sigma_i.$$

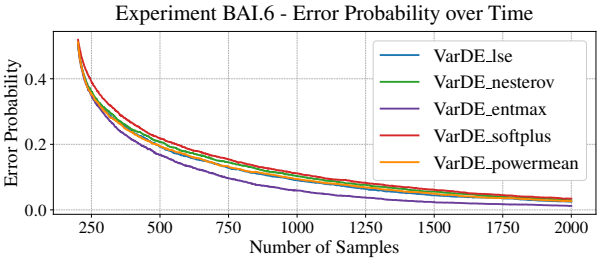
Application-level results

- **BAI:** misidentification probability decays exponentially.
- **MCTS:** wrong root-action probability decays exponentially.
- **Q-learning:** converges to an optimal policy under tabular assumptions.

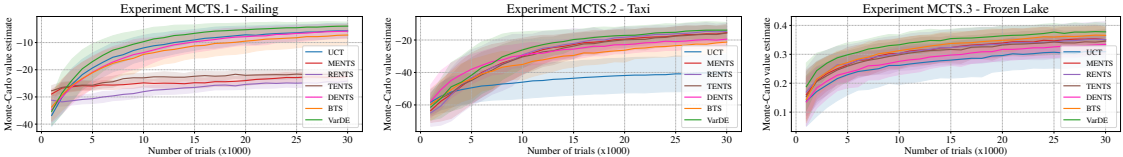
Empirical evidence: Best-arm identification

Method	BAI.1	BAI.2	BAI.3	BAI.4
Uniform	33.23	38.69	28.53	32.98
SH	29.05	29.08	15.75	19.56
SR	16.06	20.70	12.39	15.41
CR-A	17.00	21.07	9.83	11.93
CR-C	16.71	20.84	11.80	12.71
UCBE_2	15.67	21.59	12.93	19.32
UCBE_4	19.38	25.92	16.56	22.99
UCBE_8	23.54	28.59	19.25	26.09
UGapE_2	15.64	21.75	13.48	20.55
UGapE_4	20.43	26.43	17.50	23.48
UGapE_8	24.96	30.40	19.55	26.01
$\mathbf{VarDE}_{0.05}$	**12.87**	**17.27**	11.56	**11.83**
$\mathbf{VarDE}_{0.1}$	14.04	19.31	**7.34**	13.81
$\mathbf{VarDE}_{0.15}$	16.56	21.26	8.21	18.52

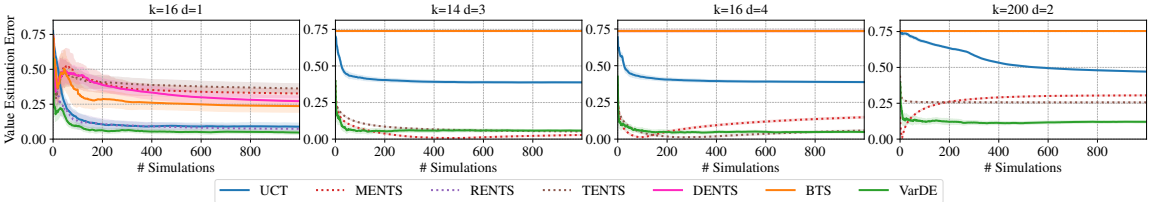




Empirical evidence: Monte Carlo tree search

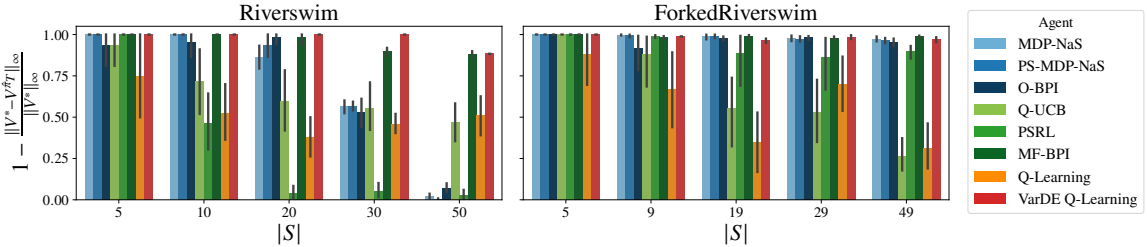


Monte Carlo value estimates of the recommended root action/policy during planning (mean $\pm$ 95% CI over runs).



Value estimation error of VarDE–MCTS and other algorithms on synthetic tree (Lower is better).

Empirical evidence: Best-policy identification



Normalized value proximity of VarDE–Q-learning against model-free and model-based methods after a fixed interaction budget T on RiverSwim and ForkedRiverSwim (mean $\pm$ 95% CI).

Takeaway

Pure exploration should optimize uncertainty of the final output.

Principle

Decision-level variance minimization.

Rule

Sample by influence $\times$ stochastic variance.

Scope

Bandits, Monte Carlo tree search, and Reinforcement Learning.

Thank you. I look forward to discussing VarDE.