ICML 2026

Academic Presentation

Active Policy Optimization for Individualized Dosing via Gradient Variance Minimization

GVALID: Gradient Variance Active Learning for Individualized Dosing

Problem
Budgeted dose policy learning

→

Insight
Regret is tied to gradient uncertainty

→

Method
Batch acquisition by variance reduction

Authors: Yi Wan | Xin Wang | Huanhuan Chen

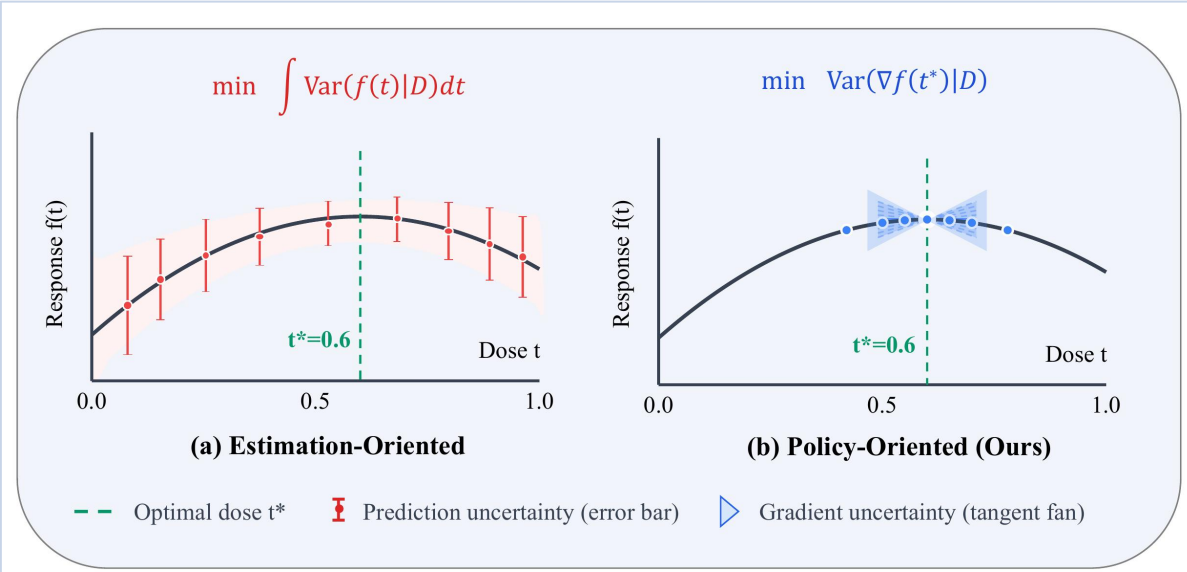
Presenter: 万仪 Yi Wan

School of Artificial Intelligence and Data Science, USTC

University of Science and Technology of China

Talk Structure

Motivation: Estimation Is Not the End Goal



Policy-oriented acquisition **focuses near the estimated optimum**, rather than **uniformly reducing uncertainty over the whole curve**.

Budget-limited interventions

- Interventions in healthcare and marketing are expensive.
- Each individual may **only receive one assigned** dose.
- Feedback is observed after a **batch**, not immediately.

Decision-focused view

- The final goal is a **high-value dosing policy**.
- Global dose-response fitting can waste samples.
- **Acquisition should be aligned with policy regret**.

Active Continuous-Dose Policy Learning

Outcome model

$$Y_x^t = f(x,t) + \varepsilon$$

→

Policy value

$$V(\pi) = E_X[f(X,\pi(X))]$$

→

Objective

$$\min R(\hat{\pi})$$

Experimental regime

Pool-based
Select from an **available individual pool**.

Batched
Update after a **cohort** of interventions.

One-shot
Each individual is queried **at most once**.

Strict budget
Only a small fraction of the pool is labeled.

$$\min R(\hat{\pi}) = V(\pi^*) - V(\hat{\pi})$$

Regret Is Controlled by Gradient Uncertainty

Posterior-mean policy

$$\hat{t}_D(x) = \arg\max_t \mu_D(x,t)$$

Policy-regret upper bound

$$\text{Policy Regret:} \quad R(D) \le C_g \cdot U_g(D)$$

$$\text{Uncertainty Bound:} \quad U_g(D) = E_X \, \mathrm{Var}[g(X, \hat{t}_D(X)) \mid D]$$

Why gradients?

- At an optimum, the dose gradient should be zero.
- Gradient uncertainty measures **uncertainty about stationarity**.
- Reducing it tightens a policy-regret bound.

Design Implication

Reduce posterior **gradient variance at the current recommended doses**, not predictive variance everywhere.

GVALID Batch Acquisition

1. Target contexts

Sample representative x from the deployment pool.

2. Current dose

Maximize the GP posterior mean over dose.

3. Gradient proxy

Use a symmetric **finite difference** around t.

4. Query batch

Greedily maximize **variance reduction**.

Finite-difference target

$$\tilde{g}_j = \frac{f(x_j, t_j + \delta) - f(x_j, t_j - \delta)}{2\delta}$$

Greedy target score

$$\Delta(q) = \frac{||\mathrm{Cov}(\tilde{g}, y_q)||_2^2}{\mathrm{Var}(y_q)}$$

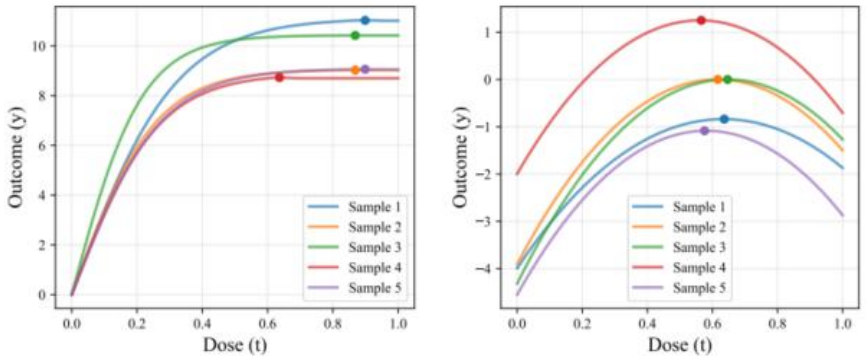
Schur-complement updates refresh candidate scores after each selected query, avoiding full covariance recomputation.

Strict Budgets, Shared GP Backbone

Evaluation protocol

Datasets

- Synthetic: HNL, CSC, SWV, STS
- Semi-synthetic: News



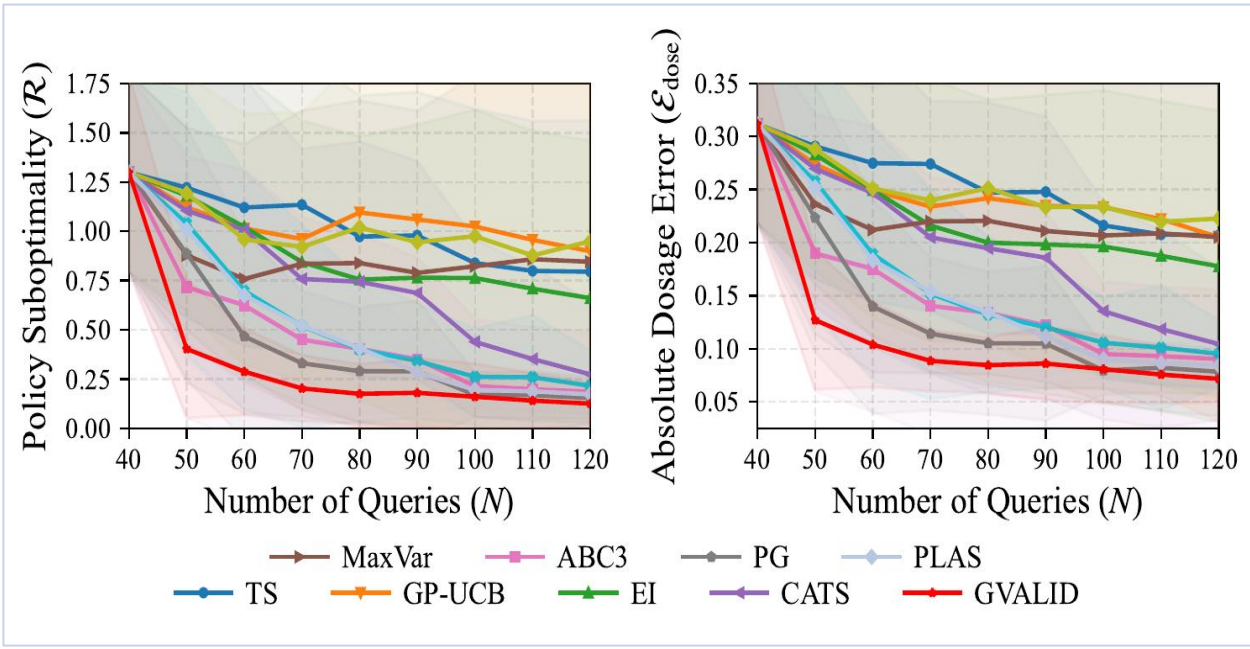
(d) STS *(e)* News

Metric and budget

- Policy suboptimality R (lower is better)
- Total budget: 20% of the training pool
- Unified GP framework across methods

Baselines

TS, GP-UCB, EI, CATS, MaxVar, ABC3, PG, SoftTopK, PLAS.



Example: News benchmark

Policy suboptimality and dosage error decrease together, supporting optimal-dose identification.

Main Results and Analysis

Representative policy suboptimality at N = 350

Dataset	GVALID	Best Baseline	Rel. Gain
HNL	**0.12**	0.17 (ABC3)	29%
CSC	**0.40**	0.50 (MaxVar)	20%
SWV	**0.29**	0.34 (PLAS)	15%
STS	**0.17**	0.17 (ABC3)	tie

Ablation at N = 300

Method	HNL	CSC	News
GVALID-RX	**0.25**	1.09	0.09
GVALID-F	**0.26**	0.77	0.30
GVALID	**0.15**	**0.48**	**0.07**

Main observation

- **Faster policy improvement** under tight budgets.
- Strongest gains when smooth dose-response assumptions are informative.
- **Competitive under multimodal or plateau-shaped responses**.

Why the components matter

- Gradient variance is more **policy-aligned** than function variance.
- **Joint context-dose acquisition** improves sample efficiency.

What this work contributes

Policy-oriented formulation	Budget-constrained individualized dosing is cast as **active policy optimization**.
Theory-guided criterion	Policy regret is connected to **posterior gradient variance** at estimated optima.
GVALID acquisition	A greedy batch rule targets **policy-relevant gradient uncertainty** under one-shot constraints.
Empirical evidence	Experiments show **improved sample efficiency** across synthetic and semi-synthetic benchmarks.

Take-home message

GVALID spends experimental budget where it most reduces uncertainty about whether the currently recommended dose is truly optimal.

Future directions

- Safety-aware dose constraints
- Scalable approximations for larger pools
- Robustness beyond smooth unimodal responses