FY2025/2025年度

Imperfect Information Learning Team

Masashi Sugiyama

不完全情報学習チーム　杉山 将

RIKEN Center for Advanced Intelligence Project

Our Vision and Social Impact

- Develop **trustworthy machine learning** methods/algorithms that can cope with imperfect training information like distribution shift, noisy labels, partial labels, and pseudo-supervision.
- Enable machine learning for real-world applications in imperfect or adversarial deployment environments such as robust image/video classification and sample-/label-efficient text classification.

Team Members

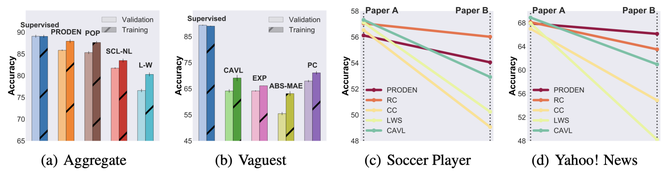
- Masashi Sugiyama (Team Director)
- Gang Niu (Senior Research Scientist)
- Takashi Ishida (Research Scientist)
- Wei Wang (Postdoc)
- Zhen-Yu Zhang (Postdoc)
- Okan Koc (Postdoc)
- Ming-kun Xie (Postdoc)
- Xinqiang Cai (Postdoc)

Weakly Supervised Learning (WSL)

Realistic Evaluation of Deep Partial-Label Learning Algorithms

Wang et al. (ICLR 2025)

- **Motivation**: Partial-label learning (PLL) methods are usually evaluated with clean-label validation data, inconsistent experimental settings, and without realistic image benchmarks, which hides how they behave in realistic scenarios.
- **Methodology**: This work introduces the **PLENCH benchmark** with principled model-selection criteria and a human-annotated partial-label CIFAR-10 dataset (PLCIFAR10), enabling standardized, realistic comparison of deep PLL algorithms.

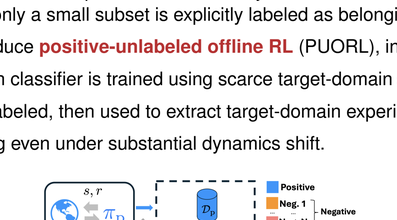


(a) Aggregate (b) Vaguest (c) Soccer Player (d) Yahoo! News

Label Distribution Learning with Biased Annotations Assisted by Multi-Label Learning

Kou et al. (IJCAI 2025)

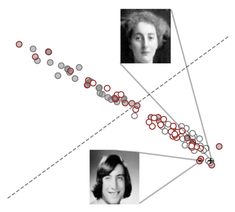
- **Motivation**: Label distribution learning must predict a soft label per instance, but real annotators often produce **biased or noisy distributions** that violate low-rank assumptions used previously.
- **Methodology**: We first convert biased soft label distributions into easier hard multi-label assignments and then recover the underlying distributions under a low-rank structure in the multi-label space, leading to more accurate and robust LDL from biased annotations.



Domain Adaptation and Entanglement: An Optimal Transport Perspective

Koc et al. (AISTATS 2025)

- **Motivation**: Unsupervised domain adaptation methods commonly align source and target feature distributions, yet existing theory does not clearly explain when such alignment guarantees robustness under distribution shift.
- **Methodology**: Using optimal transport, this work derives new generalization bounds containing an **entanglement** term—the expected Wasserstein distance between class-conditional distributions—and shows empirically how this term explains the success or failure of popular domain-adaptation algorithms.

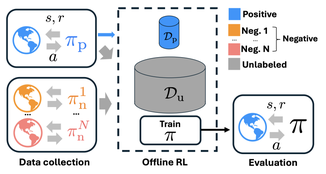


Intersection of WSL & RL

Offline Reinforcement Learning with Domain-Unlabeled Data (PUORL)

Nishimori et al. (RL Journal 2025)

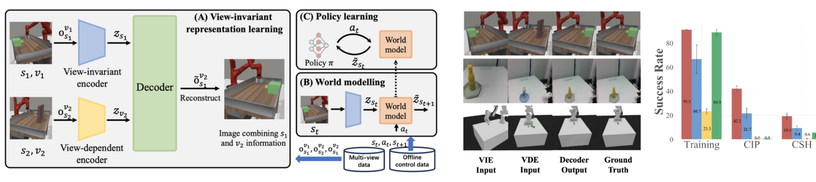
- **Motivation**: Offline RL datasets in practice often mix trajectories from multiple domains with different dynamics, while only a small subset is explicitly labeled as belonging to the target domain.
- **Methodology**: We introduce **positive-unlabeled offline RL** (PUORL), in which a PU-learning-based domain classifier is trained using scarce target-domain samples as positives and the remaining data as unlabeled, then used to extract target-domain experience that augments standard offline RL training even under substantial dynamics shift.



Learning View-invariant World Models for Visual Robotic Manipulation (ReViWo)

Pang et al. (ICLR 2025)

- **Viewpoint robustness in robotic manipulation**: Visual manipulation policies fail to generalize when **camera angles change or are disturbed** during deployment.
- **Motivation**: Existing world models focus on physical states w/ camera perspectives. We need to decouple the **state information** from the **view information** contained in the visual input for robust control.
- **Methodology**: We propose a framework that disentangles **view-invariant states** (for control) from **view-dependent features** via a factorized latent space, achieving superior robustness against viewpoint disturbances.

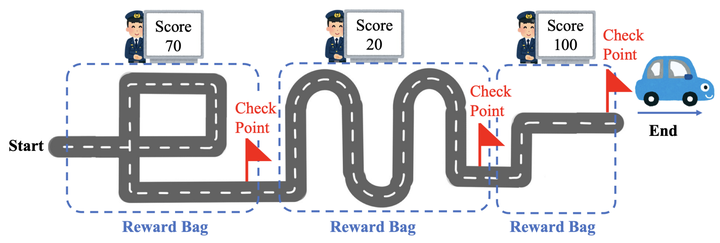


Reinforcement Learning (RL)

Reinforcement Learning from Bagged Reward

Tang et al. (TMLR 2025)

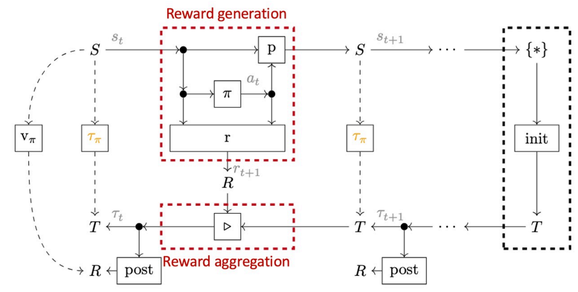
- **Motivation**: In many real-world tasks, an agent receives a single reward for a partial sequence or entire trajectory rather than step-wise rewards, making standard RL algorithms hard to apply.
- **Methodology**: This work formalizes **bagged reward** Markov decision processes and shows they can be reduced to standard MDPs via reward redistribution; it then proposes a Transformer-based reward model with bidirectional attention to allocate bagged rewards to individual steps, substantially improving learning, especially for long bags.



Recursive Reward Aggregation

Tang et al. (RL Journal 2025)

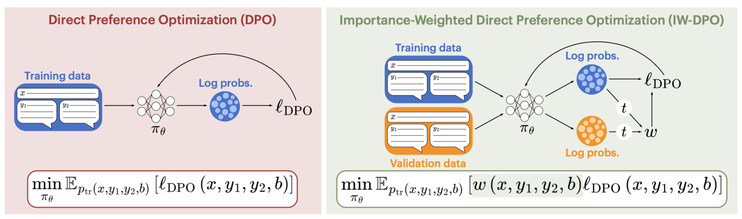
- **Motivation**: Complex objectives in RL, such as risk-sensitive or max-type criteria, cannot be expressed well as a simple discounted sum of rewards.
- **Methodology**: We introduce an algebraic framework that derives Bellman equations from **recursive reward generation** and **aggregation**, generalizes discounted cumulative return to diverse recursive objectives (e.g., discounted max, Sharpe ratio), and integrates it into value-based and actor-critic algorithms so agents can optimize diverse objectives without redesigning the underlying reward function.



Importance Weighting for Aligning LMs under Deployment Distribution Shift (IW-DPO)

Lodkaew et al. (TMLR 2025)

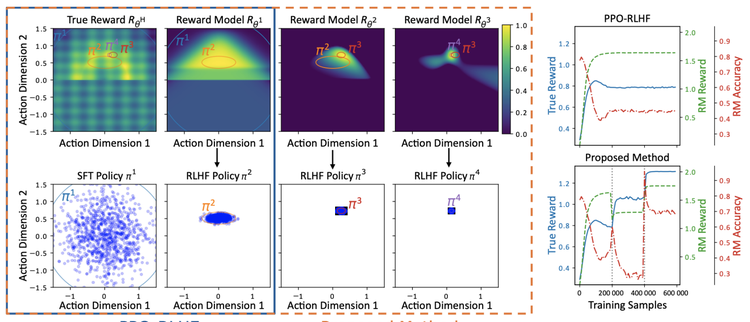
- **Motivation**: Direct Preference Optimization (DPO) aligns language models to human preferences on a fixed dataset, but performance can degrade when prompts, responses, or preference labels shift after deployment.
- **Methodology**: IW-DPO estimates **importance weights** from a small amount of post-deployment validation data using density-ratio estimation and reweights DPO's training losses so that fine-tuning directly targets the deployment distribution, improving alignment under various shift scenarios.



Off-Policy Corrected Reward Modeling for Reinforcement Learning from Human Feedback

Ackermann et al. (COLM 2025)

- **Motivation**: In RL from Human Feedback, we first train a reward model (RM) on preference data sampled from the initial policy. The RM is then used to update the policy to align better with human preferences. As the policy is changes, the RM becomes inaccurate and reward-hacking occurs.
- **Methodology**: We interpret this issue as a covariate shift problem. By retraining the RM using importance weighting, we can reuse the data sampled from the initial policy to obtain a new RM accurate on the current policy. By iterating policy updates and RM retraining, we obtain a significantly better alignment in LLM summarization and chatbot tasks.

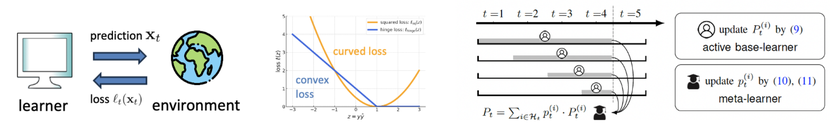


Online Learning and Bandits

Non-stationary Online Learning for Curved Losses: Improved Dynamic Regret via Mixability

Zhang et al. (ICML 2025)

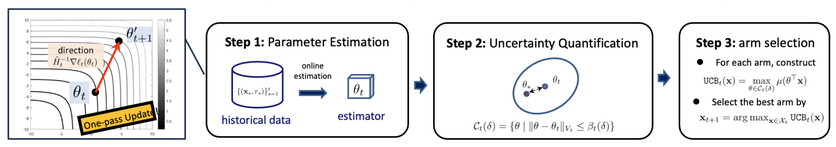
- **Non-stationary online learning**: We study how to learn with streaming data, where underlying distribution could sequentially change over each time.
- **Motivation**: We aim to accelerate learning by exploiting loss-function curvature with provable guarantees. Existing methods suffer from heavy dependence on data dimensionality.
- **Methodology**: By leveraging **mixability**, a property that reflects the curvature of the losses, we propose an online ensemble method that achieves fast-rate dynamic regret guarantees.



Generalized Linear Bandits: Almost Optimal Regret with One-Pass Update

Zhang et al. (NeurIPS 2025)

- **Generalized linear bandits**: We study decision making under uncertainty, using a generalized linear model to capture the categorical or count-based feedback common in many applications.
- **Motivation**: Their non-linear feedback makes a trade-off: existing methods either incur high per-round costs for optimal regret or rely on constant-time updates that sacrifice statistical efficiency.
- **Methodology**: We propose a jointly efficient method based on **online mirror descent** that retains near-optimal statistical efficiency while achieving constant computation and storage costs.



Theory of Parallel Optimization and Sampling

Parallel Simulation for Log-concave Sampling and Score-based Diffusion Models

Zhou & Sugiyama (ICML 2025)

- **Motivation**: Sampling from high-dimensional log-concave distributions and score-based diffusion models can be bottlenecked by the number of sequential update rounds required by existing algorithms.
- **Theory**: Building on the observation that both strongly log-concave sampling and sampling for score-based diffusion models rely on simulating closely related dynamics, each with a smooth drift and the same time length, we develop a parallel sampling scheme using parallel simulation techniques. This approach reduces the adaptive-complexity dependence on the dimension from $\mathcal{O}(\log^2 d)$ to $\mathcal{O}(\log d)$.

Table 1. Comparison with existing parallel methods for strongly log-concave sampling. The symbol ♠ represents that the results hold under a weaker condition, the log-Sobolev inequality.

Work dynamics	Measure	Adaptive Complexity	Space Complexity
(Shen & Lee, 2019, Theorem 4) underdamped Langevin diffusion	W_2	$\widetilde{\mathcal{O}}(\log^2(\frac{d}{\varepsilon}))$	$\widetilde{\mathcal{O}}(\frac{d}{\varepsilon})$
(Vu & Dalalyan, 2024, Corollary 2) underdamped Langevin diffusion	W_2	$\widetilde{\mathcal{O}}(\log^2(\frac{d}{\varepsilon}))$	$\widetilde{\mathcal{O}}(\frac{d}{\varepsilon})$
(Anari et al., 2024, Theorem 15) underdamped Langevin diffusion	TV	$\widetilde{\mathcal{O}}(\log^2(\frac{d}{\varepsilon}))^{\spadesuit}$	$\widetilde{\mathcal{O}}(\frac{d}{\varepsilon})$
(Anari et al., 2024, Theorem 13) overdamped Langevin diffusion	KL	$\widetilde{\mathcal{O}}(\log^2(\frac{d}{\varepsilon}))^{\spadesuit}$	$\widetilde{\mathcal{O}}(\frac{d}{\varepsilon})$
Theorem 4.2 overdamped Langevin diffusion	KL	$\widetilde{\mathcal{O}}(\log(\frac{d}{\varepsilon}))$	$\widetilde{\mathcal{O}}(\frac{d}{\varepsilon})$

Table 2. Comparison with existing parallel methods for sampling for diffusion models.

Works Implementation	Measure	Adaptive Complexity	Space Complexity
(Chen et al., 2024, Theorem 3.5) ODE / Picard method	TV	$\widetilde{\mathcal{O}}(\log^2(\frac{d}{\varepsilon^2}))$	$\widetilde{\mathcal{O}}(\frac{d^{2/2}}{\varepsilon^2})$
(Gupta et al., 2025, Theorem B.13) ODE / Parallel midpoint method	TV	$\widetilde{\mathcal{O}}(\log^2(\frac{d}{\varepsilon^2}))$	$\widetilde{\mathcal{O}}(\frac{d^{2/2}}{\varepsilon^2})$
(Chen et al., 2024, Theorem 3.3) SDE / Picard method	KL	$\widetilde{\mathcal{O}}(\log^2(\frac{d}{\varepsilon^2}))$	$\widetilde{\mathcal{O}}(\frac{d^2}{\varepsilon^2})$
Theorem 5.2 SDE / Parallel Picard method	KL	$\widetilde{\mathcal{O}}(\log(\frac{d}{\varepsilon^2}))$	$\widetilde{\mathcal{O}}(\frac{d^2}{\varepsilon^2})$

The Adaptive Complexity of Finding a Stationary Point

Zhou et al. (COLT 2025)

- **Motivation**: For non-convex optimization, it is unclear how much benefit massive parallelization can bring compared with standard sequential dimensional-free gradient-based methods when seeking a approximate stationary point.
- **Theory**: We establish **tight lower bounds** on the number of parallel rounds required for high-dimensional smooth functions and show that classic methods such as gradient descent and p-order adaptive regularization are already adaptively optimal. In the constant-dimensional regime, we propose a constant-round algorithm to trap the gradient flow together with **tight** lower bounds establishing the adaptive optimality of gradient-flow trapping-type methods.

Problem setting		Adaptive complexity	Queries per iteration
(constant) $d = \Theta(1)$ $(d \geq 2)$	Upper bounds	$k = \Theta(1)$ (Theorem 10)	$\mathcal{O}\left(\varepsilon^{-\frac{(d+1)^k}{(2d)^k-(d+1)^k}\cdot\frac{d-1}{2}}\right)$
	Lower bounds	$k = \Theta(1)$ (Theorem 11)	$\widetilde{\Omega}\left(\varepsilon^{-\frac{d^k}{d^k-1}\cdot\frac{d-1}{2}}\right)$
	Upper bounds	$\Theta\left(\log\left(\frac{1}{\varepsilon}\right)\right)$ (Hollender and Zampetakis, 2023)	$\mathcal{O}\left(\varepsilon^{-(d-1)/2}\right)$
	Lower bounds	$\Theta(\log(\frac{1}{\varepsilon}))$ (Theorem 11)	$\widetilde{\Omega}\left(\varepsilon^{-(d-1)/2}\right)$
(high-dim.) $d = \widetilde{\Omega}\left(\varepsilon^{-\frac{2+2p}{p}}\right)$	Upper bounds	$\mathcal{O}\left(\varepsilon^{-\frac{1+p}{p}}\right)$ (Birgin et al., 2017)	1
	Lower bounds	$\Omega\left(\varepsilon^{-\frac{1+p}{p}}\right)$ (Carmon et al., 2020)	1
	Lower bounds	$\Omega\left(\varepsilon^{-\frac{1+p}{p}}\right)$ (Theorem 4)	$\mathsf{poly}(d)$

The Adaptive Complexity of Minimizing Relative Fisher Information

Zhou & Sugiyama (NeurIPS 2025)

- **Motivation**: Non-log-concave sampling remains challenging, particularly for multimodal case, and recent work has introduced an analogue of non-convex stationary-point analysis for minimizing relative Fisher information. However, existing algorithms in this framework are highly sequential with limited understanding of the adaptive complexity.
- **Theory**: To obtain a relative Fisher information of at most ε^2 from the target distribution, we propose a parallel sampling algorithm that reduces the adaptive complexity from $\mathcal{O}(d^2/\varepsilon^4)$ to $\mathcal{O}(d/\varepsilon^2)$ with dimension d and accuracy ε, and proves **matching lower bounds** in a low-accuracy regime.

Table 1: Comparisons of our lower bounds and upper bounds. Here, $\widetilde{\Omega}$ and $\widetilde{\mathcal{O}}$ omit logarithmic factors. K_0 denotes the initial KL divergence, defined as $K_0 = \mathrm{KL}(\mu_0\|\pi)$, where the initial point is drawn from the distribution μ_0.

Works	Adaptive Complexity	Queries per Iteration
Sequential averaged Langevin Monte Carlo [BCE+22, Theorem 2]	$\mathcal{O}(\frac{L^2 d K_0}{\varepsilon^4})$	1
Parallelized averaged Langevin Monte Carlo Theorem 3.1	$\mathcal{O}(\frac{L K_0}{\varepsilon^2} + \log(\frac{Ld}{\varepsilon^2}))$	$\widetilde{\mathcal{O}}(\frac{L^2 d K_0}{\varepsilon^4})$
Lower bound for $\varepsilon = \sqrt{Ld}$ [CGLL23, Theorem 9]	$\Omega(\frac{K_0}{d})$	1
Lower bound for $\varepsilon = \sqrt{Ld}$ Theorem 4.1	$\Omega(\frac{K_0}{d})$	$\mathsf{poly}(d)$

References

Papers in This Poster

[WSL1] W. Wang, D.-D. Wu, J. Wang, G. Niu, M.-L. Zhang, M. Sugiyama. *Realistic Evaluation of Deep Partial-Label Learning Algorithms*. ICLR 2025.

[WSL2] Z. Kou, S. Qin, H. Wang, J. Wang, M. Xie, S. Chen, Y. Jia, T. Liu, M. Sugiyama, X. Geng. *Label Distribution Learning with Biased Annotations Assisted by Multi-Label Learning*. IJCAI 2025.

[WSL3] O. Koc, A. Soen, C.-K. Chiang, M. Sugiyama. *Domain Adaptation and Entanglement: An Optimal Transport Perspective*. AISTATS 2025.

[RL1] Y. Tang, X.-Q. Cai, Y.-X. Ding, Q. Wu, G. Liu, M. Sugiyama. *Reinforcement Learning from Bagged Reward*. TMLR 2025.

[RL2] Y. Tang, Y. Zhang, J. Ackermann, Y.-J. Zhang, S. Nishimori, M. Sugiyama. *Recursive Reward Aggregation*. Reinforcement Learning Journal, 2025.

[RL3] T. Lodkaew, T. Fang, T. Ishida, M. Sugiyama. *Importance Weighting for Aligning Language Models under Deployment Distribution Shift*. TMLR 2025.

[RL4] J. Ackermann, T. Ishida, M. Sugiyama. *Off-Policy Corrected Reward Modeling for Reinforcement Learning from Human Feedback*. COLM 2025.

[RL&WSL1] S. Nishimori, X.-Q. Cai, J. Ackermann, M. Sugiyama. *Offline Reinforcement Learning with Domain-Unlabeled Data*. Reinforcement Learning Journal, 2025.

[RL&WSL2] J.-C. Pang, N. Tang, K. Li, Y. Tang, X.-Q. Cai, Z.-Y. Zhang, G. Niu, M. Sugiyama, Y. Yu. *Learning View-invariant World Models for Visual Robotic Manipulation*. ICLR 2025.

[OL1] Y.-J. Zhang, P. Zhao, M. Sugiyama. *Non-stationary Online Learning for Curved Losses: Improved Dynamic Regret via Mixability*. ICML 2025.

[OL2] Y.-J. Zhang, S.-A. Xu, P. Zhao, M. Sugiyama. *Generalized Linear Bandits: Almost Optimal Regret with One-Pass Update*. NeurIPS 2025.

[Th1] H. Zhou, M. Sugiyama. *Parallel Simulation for Log-concave Sampling and Score-based Diffusion Models*. ICML 2025.

[Th2] H. Zhou, A. Han, A. Takeda, M. Sugiyama. *The Adaptive Complexity of Finding a Stationary Point*. COLT 2025.

[Th3] H. Zhou, M. Sugiyama. *The Adaptive Complexity of Minimizing Relative Fisher Information*. NeurIPS 2025.