# Semi-Supervised Learning with Noisy Proxy Covariates: Generalization Bounds and Distribution Regression

Kwangho Kim[1] Jisu Kim[2]

[1]Korea University

[2]Seoul National University

ICML 2026, Seoul, South Korea

# Background: Data Generation Process

**Traditional Regression Model**

- We assume the standard non-parametric regression setup:

$$Y_i = f(X_i) + \mu_i, \quad \mathbb{E}[\mu_i | X_i] = 0$$

  - $X_i \in \mathcal{X}$: True latent (unobserved) covariates.
  - $Y_i \in \mathbb{R}$: Observed labels.
  - $\mu_i$: Mean-zero random noise.

**The Challenge in Our Setting: Noisy Proxies**

- In practice, we **do not observe** the true $X_i$.
- Instead, we only observe a contaminated proxy $\hat{X}_i$:

$$\hat{X}_i = X_i + \epsilon_i \quad \text{(Measurement Error / Noise)}$$

- **Goal:** Predict $Y$ from $\hat{X}$ by leveraging a massive amount of unlabeled proxies.

# Problem Setup: SSL with Error-in-Variables

- **Labeled Pairs ($n$):** $\{(\hat{X}_i, Y_i)\}_{i=1}^{n}$ (Both proxy and label are available)
- **Unlabeled Proxies ($N$):** $\{\hat{X}_i\}_{i=n+1}^{N}$ ($N \gg n$, Only noisy proxies)
- **Hypothesis Space:** Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\kappa$ with Mercer kernel $\kappa$.

**Target Latent Predictor:**

- We aim to approximate the best-in-class latent predictor $g$:

$$g \in \arg\min_{h \in \mathcal{H}_\kappa} \mathbb{E}\left[(Y - h(X))^2\right]$$

# The Proposed Two-Stage Framework

**Stage 1: Spectral Feature Extraction**

- Use all $N$ proxies to learn data geometry via empirical kernel integral operator:

$$h \mapsto \widehat{\mathscr{L}_N}(h)(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \kappa(\hat{X}_i, \cdot)h(\hat{X}_i).$$

- Extract leading $s$ eigenfunctions $\{\hat{\phi}_j\}_{j=1}^{s}$ to form: $\widehat{\Phi}_s(x) = (\hat{\phi}_1(x), \dots, \hat{\phi}_s(x))^\top$.

**Stage 2: Ridge Regression**

- Fit on labeled pairs ($n$) using learned data-dependent features:

$$\hat{\gamma} \in \arg\min_{\gamma} \left\{ \frac{1}{n}\sum_{i=1}^{n} \Big( Y_i - \langle \gamma, \widehat{\Phi}_s(\hat{X}_i) \rangle \Big)^2 + \xi \left\| \gamma \right\|_2^2 \right\}.$$

- Final Predictor: $\hat{g}(\cdot) = \langle \hat{\gamma}, \widehat{\Phi}_s(\cdot) \rangle$.

# Core Theory: Excess Risk Bound

Theorem

*Under polynomial decay $\lambda_j \asymp j^{-q}$, where $\lambda_j$ is the eigenvalue for population kernel integral operator, the excess risk satisfies:*

$$\mathbb{E}\left[(\hat{g}(X) - g(X))^2\right] = \mathcal{O}_P\left(s^{-q} + \frac{1}{\lambda_s^2}\left(\frac{1}{N} + C_{\beta',\theta}\right) + \frac{s}{n} + \frac{\xi}{\xi + \lambda_s^2}\right),$$

*where $C_{\beta',\theta}$ is a proxy perturbation constant.*

- Oracle Feature Rate: If $N \to \infty$ and proxy perturbation $C_{\beta',\theta} \to 0$, we recover the oracle feature rate $\mathcal{O}_P(n^{-q/(q+1)})$.
- Unlabeled data ($N$) explicitly drives down the spectral estimation error.

# Special Case: Distribution Regression

Our framework directly specializes to **Distribution Regression**, where proxies $\widehat{P}$ are empirical distributions sampled from bags of size $m$.

| **Component** | **Generic SSL with Proxy** | **Distribution Regression** |
|---|---|---|
| Latent Covariate | $X$ | Distribution $P$ |
| Noisy Proxy | $\hat{X}$ | Empirical Bag $\widehat{P}$ |
| Perturbation Floor | $C_{\beta',\theta}$ | Bag size envelope $a_m \asymp m^{-\beta/(2+d)}$ |

- Fast rates are fully recovered as the bag size $m \to \infty$.

# Summary

**Key Takeaways:**

1. Unified framework for SSL with noisy proxy covariates.
2. First explicit finite-sample bounds showing variance reduction from unlabeled proxies.
3. Distribution regression generalized as a special case.

**Thank You!**

Paper published in PMLR Vol 306, 2026.