# Monotonic Variational Gaussian Process for Efficient Data Collection

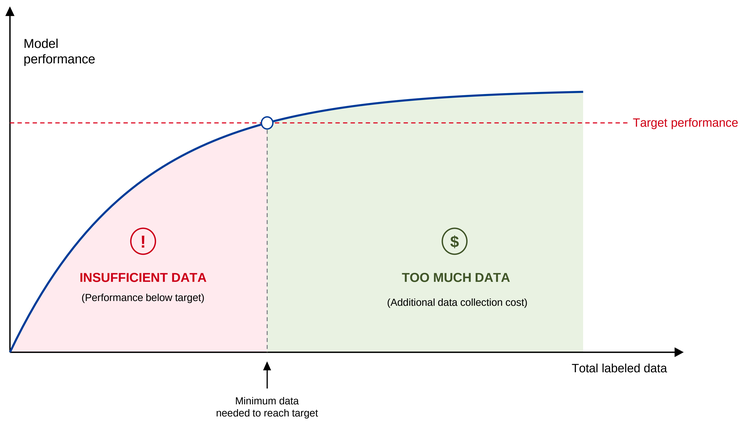
**Donghyun Lee[1], Young Myoung Ko[1*]**

[1] Pohang University of Science and Technology (POSTECH).

*: Corresponding author

## Introduction

Training deep learning models requires large datasets, but data collection is costly.




The data collection problem is to decide how much data to collect from each source to reach a target performance at minimal cost.

- At each stage $t$, we choose a cumulative labeled-count vector $\boldsymbol{n}_t$ from $K$ sources.
- More data from any source should not worsen performance : $\boldsymbol{n}_0 \leq \boldsymbol{n}_1 \leq \cdots \leq \boldsymbol{n}_T$

Prior work formalizes this as a cost minimization problem:

$$\mathcal{L}(\mathbf{n}_1, \ldots, \mathbf{n}_T) = \sum_{t=1}^{T} \mathbf{c}^\top (\mathbf{n}_t - \mathbf{n}_{t-1}) \left(1 - F(\mathbf{n}_{t-1})\right) + \gamma \left(1 - F(\mathbf{n}_T)\right).$$

Where $F(\boldsymbol{n}) = P(V(\boldsymbol{n}) \geq V^*)$ is estimated via a parametric learning curve.

Challenge:

- Parametric forms are **valid only over restricted data ranges** and extrapolate poorly.
- In **multi-source settings**, each source exhibits different and unknown scaling behavior, making a single parametric form even more restrictive.
- The failure probability penalty induces **vanishing gradients**, providing no optimization signal when performance is far below the target.

## Contribution

- We propose a **monotonic variational GP** with virtual-derivative factors, enabling tractable posterior inference without requiring derivative observations.
- We develop an **expected shortfall objective** for target-driven data collection, with theoretical guarantees of non-vanishing gradient signals.

## Method

Monotonic Variational Gaussian Process

1. Virtual derivative constraint

$$g_{j,k} := \frac{\partial f}{\partial x_k}(\tilde{\mathbf{x}}_j), \qquad p(s_{j,k} = 1 \mid g_{j,k}) = \Phi(g_{j,k})$$

Virtual binary variables $s_{j,k} = 1$ act as soft monotonicity constraints via a probit likelihood – no derivative observations required.

2. Variational Inference

$$\boldsymbol{u} = (\boldsymbol{u}_f, \boldsymbol{u}_g)^\top, \qquad q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$$

Joint inducing variables couple function values and coordinate-wise derivatives.

$$\begin{aligned}\mathrm{ELBO} = &\sum_{i=1}^{N} \mathbb{E}_{q(f(\mathbf{x}_i))} \left[\log \mathcal{N}\left(y_i \mid f(\mathbf{x}_i), \sigma^2\right)\right] \\ &+ \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbb{E}_{q(g_{j,k})} \left[\log p(s_{j,k} \mid g_{j,k})\right] \\ &- \mathrm{KL}\left(q(\mathbf{u}) \,\|\, p(\mathbf{u})\right)\end{aligned}$$

Expected Shortfall Objective

The full multi-stage objective replaces the failure probability penalty with expected shortfall $S(\boldsymbol{n})$:

$$\mathcal{L} = \sum_{t=1}^{T} \mathbf{c}_t^\top (\mathbf{n}_t - \mathbf{n}_{t-1}) R(\mathbf{n}_{t-1}) + \gamma S(\mathbf{n}_T)$$

$$S(\mathbf{n}) = (V^* - \hat{\mu})\Phi(\alpha) + \hat{\sigma}\varphi(\alpha)$$

**Algorithm** MOVE

**Input:** initial labeled set $\mathcal{D}_0$, target $V^*$, cost $\mathbf{c} \in \mathbb{R}_+^K$, Horizon $T$
**repeat for** each stage $t = 1, \ldots, T$
  Maximize ELBO to fit monotonic variational GP
  Obtain $\hat{\mu}(\cdot)$, $\hat{\sigma}(\cdot)$ from $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$
  Minimize $\mathcal{L}$ over $(\mathbf{n}_1, \ldots, \mathbf{n}_T)$ via projected SGD
  Collect labels; train model; observe $V(\mathbf{n}_t)$
**until** $t = T$ **or** $V^*$ achieved
**Output:** plan $(\mathbf{n}_1, \ldots, \mathbf{n}_T)$

## Theoretical Analysis

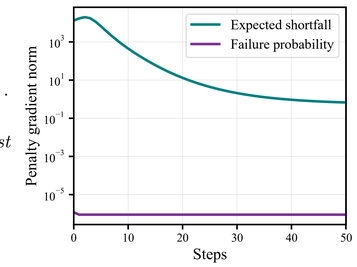
**Theorem 3.4.** *Suppose Assumption 3.2 and 3.3 hold. Then*

$$\nabla R(\mathbf{n}) = \varphi(\alpha(\mathbf{n})) \ \nabla\alpha(\mathbf{n}),$$

*and there exist constants $C_1, C_2 > 0$ such that*

$$\|\nabla R(\mathbf{n})\| \leq C_1 \varphi(\alpha(\mathbf{n})) \ (1 + C_2|\alpha(\mathbf{n})|$$
$$= \frac{C_1}{\sqrt{2\pi}}(1 + C_2|\alpha(\mathbf{n})|) \ \exp(-\alpha(\mathbf{n})^2/2) .$$

*Consequently, $\|\nabla R(\mathbf{n})\|$ decays at least exponentially fast in $|\alpha(\mathbf{n})|$ as $|\alpha(\mathbf{n})| \to \infty$.*




**Theorem 3.5.** *Suppose Assumption 3.2 holds. Then*

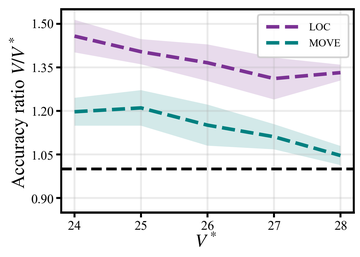
$$\nabla S(\mathbf{n}) = -\Phi(\alpha(\mathbf{n})) \ \nabla\mu(\mathbf{n}) + \varphi(\alpha(\mathbf{n})) \ \nabla\sigma(\mathbf{n}),$$

*Consequently, $\nabla S(\mathbf{n})$ approaches $-\nabla\mu(\mathbf{n})$ as $\alpha(\mathbf{n}) \to \infty$.*

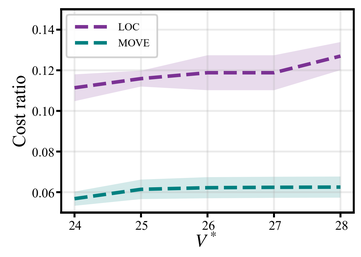
## Experiments

| Dataset | Task | Metric | Full data set size |
|---|---|---|---|
| CIFAR-10 (Krizhevsky, Hinton, et al. 2009) | Classification | Top-1 Accuracy | 50,000 |
| CIFAR-100 (Krizhevsky, Hinton, et al. 2009) | Classification | Top-1 Accuracy | 50,000 |
| ImageNet (Deng et al. 2009) | Classification | Top-1 Accuracy | 1,281,167 |
| VOC-Seg (Hariharan et al. 2011; Everingham et al. 2015) | Semantic Segmentation | Mean IoU | 10,582 |
| VOC-Det (Everingham et al. 2015) | Object Detection | Mean AP | 16,551 |

| Task | Dataset | $T$ | LOC | | ExactGP | | MOVE | |
|---|---|---|---|---|---|---|---|---|
| | | | Acc. ratio | Cost ratio | Acc. ratio | Cost ratio | Acc. ratio | Cost ratio |
| Class. | CIFAR-10 | 3 | 0.04 | 0.0568 | 0.04 | 0.0640 | 0.01 | **0.0433** |
| | | 5 | 0.03 | 0.0484 | 0.03 | 0.0563 | 0.02 | **0.0423** |
| | CIFAR-100 | 3 | 0.19 | 0.0633 | 0.19 | 0.0638 | 0.07 | **0.0404** |
| | | 5 | 0.16 | 0.0525 | 0.20 | 0.0592 | 0.07 | **0.0334** |
| | ImageNet | 3 | -0.03 | 0.0016 | 0.02 | 0.0144 | 0.02 | **0.0117** |
| | | 5 | -0.04 | 0.0034 | 0.02 | 0.0107 | 0.01 | **0.0099** |
| Seg. | VOC-Seg | 3 | 0.06 | 0.1516 | 0.08 | 0.1647 | 0.00 | **0.1045** |
| | | 5 | 0.05 | 0.1308 | 0.07 | 0.1465 | 0.01 | **0.1041** |
| Det. | VOC-Det | 3 | 0.54 | 0.1538 | 0.18 | **0.0587** | 0.19 | 0.0633 |
| | | 5 | 0.40 | 0.1159 | 0.26 | 0.0785 | 0.21 | **0.0614** |




CIFAR-10(Cost)




VOC-Det(Cost)