Marginal Distribution Adaptation for Discrete Sets via Module-Oriented Divergence Minimization

Presented by Hanjun Dai,

Joint work with Sherry Yang, Emily Xue, Dale Schuurmans, Bo Dai

Google Research

Background

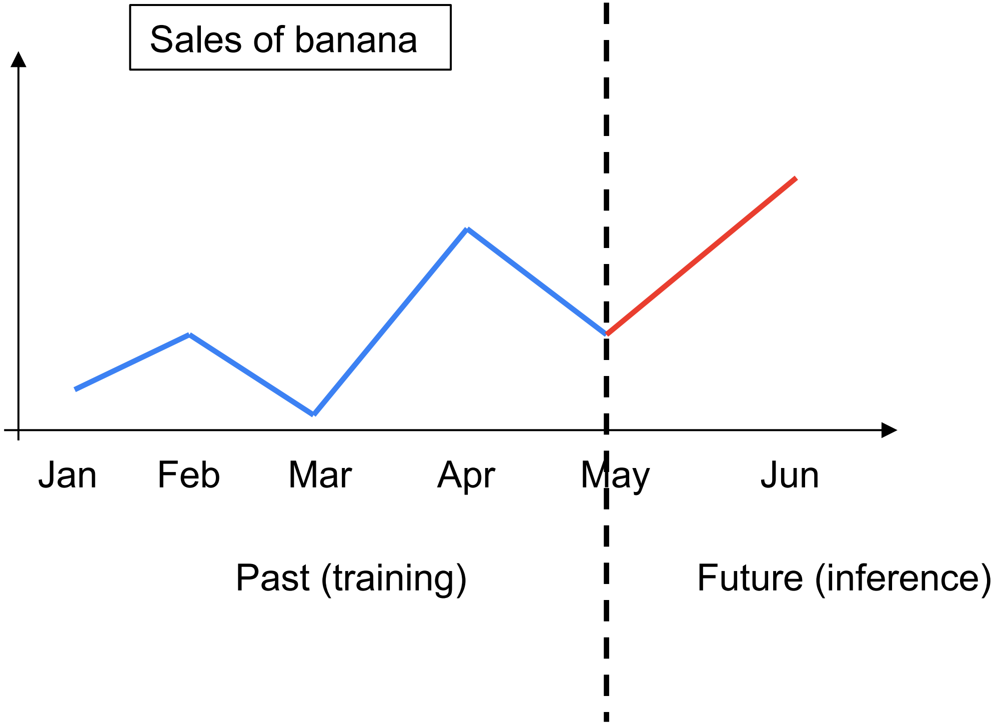
Generative modeling of discrete sets

Cart modeling

Background

Generative modeling of discrete sets

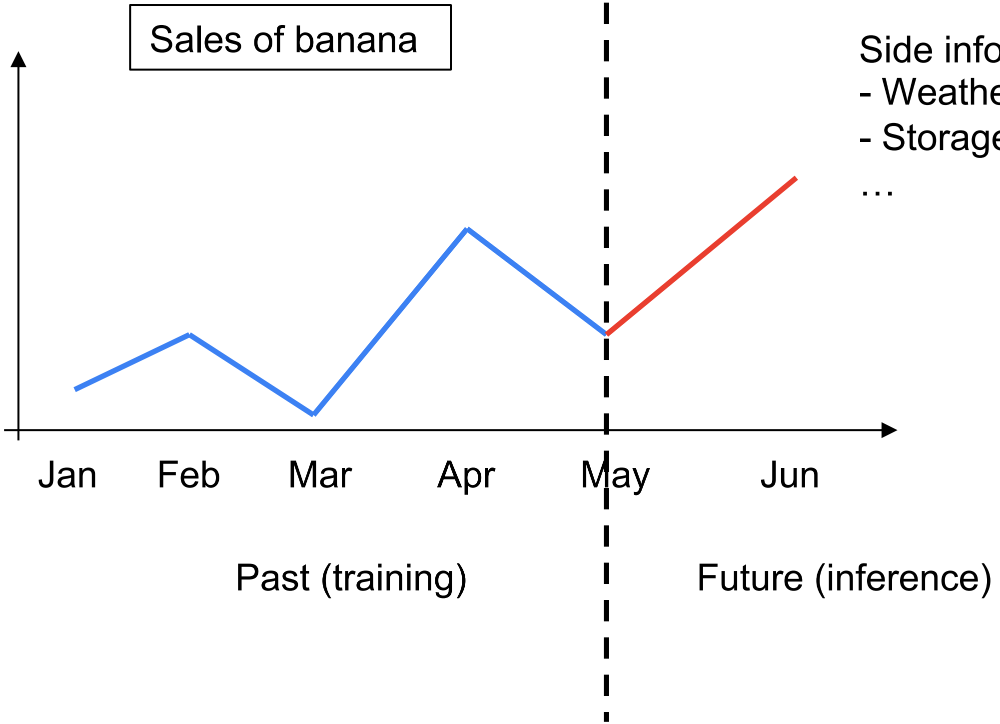
Cart modeling



Background

Generative modeling of discrete sets

Cart modeling



Side information
- Weather
- Storage
…

Research Question

How can we efficiently align an existing generative model to match target marginal specifications, while preserving previously learned correlations between elements?

Problem formulation

$$\min_{q \in \mathcal{H}} \quad KL(q||p)$$

$$\text{s.t.} \quad |\mathbb{E}_{S \sim q}[\mathbb{I}(e_i \in S)] - t_i| \leqslant \epsilon \quad \forall (e_i, t_i) \in C.$$

Problem formulation

Adapted model — Existing generative model

$$\min_{q \in \mathcal{H}} \quad KL\left(q \| p\right)$$

$$\text{s.t.} \quad \left|\mathbb{E}_{S \sim q}\left[\mathbb{I}\left(e_i \in S\right)\right] - t_i\right| \leqslant \epsilon \quad \forall\left(e_i, t_i\right) \in C.$$

Google Research

Problem formulation

$$\min_{q \in \mathcal{H}} \quad KL\left(q || p\right)$$

$$\text{s.t.} \quad \left|\mathbb{E}_{S \sim q}\left[\mathbb{I}\left(e_i \in S\right)\right] - t_i\right| \leqslant \epsilon \quad \forall \left(e_i, t_i\right) \in C.$$

- Adapted model: q
- Existing generative model: p
- Target marginal: t_i
- Discrete set sampled from q: S
- Element (e.g., apple) in the set: e_i
- All marginal specifications: C

Instantiations

$$q \leftarrow p$$

- Same distribution family
- Reusing part of the model p

Instantiations

$$q \leftarrow p$$

- Same distribution family
- Reusing part of the model p

⟹

- Minimize the # updated parameters
- Improve sample efficiency

Instantiations

Derivation of marginal distribution

Constrained optimization

- Latent variable models
- Autoregressive models
- Energy based models

Instantiations

Derivation of marginal distribution

Constrained optimization

- **Latent variable models**
- Autoregressive models
- Energy based models

Google Research

Latent variable model for sets

$$p(B) = \int_\theta p(\theta) \prod_{i=1}^{|X|} p(B_i|\theta)$$

Google Research

Latent variable model for sets

$$p(B) = \int_\theta p(\theta) \prod_{i=1}^{|X|} p(B_i|\theta)$$

Marginal distribution:

$$\begin{aligned} p(B_i) &= \sum_{\tilde{B}\in\{0,1\}^{|X|}, \tilde{B}_i = B_i} \int_\theta p(\theta) \prod_{j=1}^{|X|} p(\tilde{B}_j|\theta) \\ &= \int_\theta p(\theta) p(B_i|\theta) \left(\sum_{\tilde{B}} \prod_{j\neq i} p(\tilde{B}_j|\theta) \right) \\ &= \int_\theta p(\theta) p(B_i|\theta) \end{aligned}$$

Google Research

Latent variable model for sets

Existing learned model: $p(B) = \int_\theta p(\theta) \prod_{i=1}^{|X|} p(B_i|\theta)$

Adapted model:
$$q(B) = \int_\theta q(\theta) \prod_{i=1}^{|X|} p(B_i|\theta)$$

$$\min_{q(\theta)} \quad KL\left(q(\theta)||p(\theta)\right)$$
$$\text{s.t.} \quad \left|\left|\mathbb{E}_{\theta\sim q(\theta)}\left[p(B_{e_i}|\theta)\right] - t_i\right|\right|_2 \leqslant \epsilon, \ \forall(e_i, t_i) \in C.$$

Google Research

Instantiations

Derivation of marginal distribution

Constrained optimization

- Latent variable models
- **Autoregressive models**
- Energy based models

Google Research

Autoregressive model for sets

$$p(S|L) = \prod_{i=1}^{L} p(s_i|s_{<i}, L)$$

Order invariant assumption for sets:

$$p(S^{\pi}|L) = \prod_{i=1}^{L} p(s_{\pi_i}|s_{<\pi_i}, L) = p(S^{\pi'}|L)$$

Autoregressive model for sets

$$p(S|L) = \prod_{i=1}^{L} p(s_i|s_{<i}, L)$$

Marginal distribution:

$$\begin{aligned} p(x) &= \sum_{L=1}^{|X|} p(L) \sum_{S:|S|=L} p(x \in S|L) \\ &= \sum_{L=1}^{|X|} p(L) \sum_{S:|S|=L} p(s_1 = x|L) \times L \end{aligned}$$

Autoregressive model for sets

Existing learned model: $p(S|L) = \prod_{i=1}^{L} p(s_i|s_{<i}, L)$

Adapted model: $$q(S) = p(|S|)q_1(s_1||S|)\prod_{i=2}^{|S|} p(s_i|s_{<i}, |S|)$$

$$\min_{q_1} \quad \mathbb{E}_{L\sim p(L)} KL\left(q_1(\cdot|L)||p_1(\cdot|L)\right)$$

$$\text{s.t.} \quad \|q(e_i) - t_i\|_2 \leqslant \epsilon, \forall (e_i, t_i) \in C$$

Google Research

Instantiations

Derivation of marginal distribution

Constrained optimization

- Latent variable models
- Autoregressive models
- **Energy based models**

Energy based model for sets

$$p_f(B) = \frac{\exp(f(B))}{Z_p}, \ Z_f = \sum_{B \in \{0,1\}^{|X|}} \exp(f(B))$$

Primal optimization problem:

$$\min_{q \in \mathcal{P}} KL\left(q || p_f\right) \quad \text{s.t.} \ \left\|\mathbb{E}_q\left[\phi\left(B\right)\right] - c\right\|_2 \leqslant \epsilon,$$

Equivalent dual form:

$$\max_{w} w^\top c - \log \sum_{B} \exp(w^\top \phi(B) + f(B)) - \epsilon \left\|w\right\|_2$$

Google Research

Experiments

Pairwise F1

$$\text{Precision} = \frac{\sum_{x,y} \min\left\{c2(x, y; \mathcal{D}_{gen}), c2(x, y; \mathcal{D}_{tgt})\right\}}{c2(\mathcal{D}_{gen})}$$

and the recall as:

$$\text{Recall} = \frac{\sum_{x,y} \min\left\{c2(x, y; \mathcal{D}_{gen}), c2(x, y; \mathcal{D}_{tgt})\right\}}{c2(\mathcal{D}_{tgt})}$$

Marginal RMSE

$$\sqrt{\frac{1}{|C|} \sum_{(e_i, t_i) \in C} \left(t_i - \frac{\sum_{S \in \mathcal{D}} \mathbb{I}(e_i \in S)}{|\mathcal{D}|}\right)}$$

Google Research

Real-world experiments

Table 1. Real-world dataset statistics.

Dataset	$\lvert\mathcal{D}_{src}\rvert$	$\lvert\mathcal{D}_{tgt}\rvert$	$\lvert X\rvert$	MaxSetSize
Groceries	8,851	984	169	32
Market-Basket	13,466	1,497	167	10
MIMIC3	53,030	5,893	1,070	39
MIMIC3-sec	53,030	5,893	19	16
Instacart	2,963,177	119,533	1,000	79

Google Research

Real-world experiments

Lower marginal RMSE after adaptation

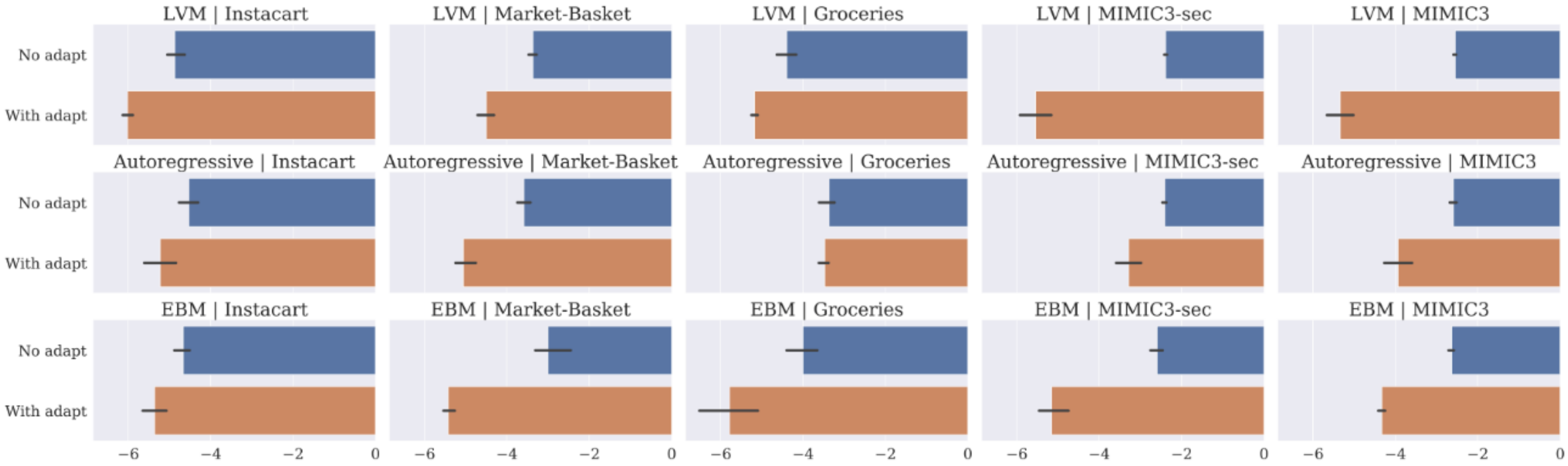


Figure 4. Marginal log-RMSE for models before and after marginal adaptations on real-world datasets.

Google Research

Real-world experiments

Similar pairwise F1 ---- maintains the correlations between items

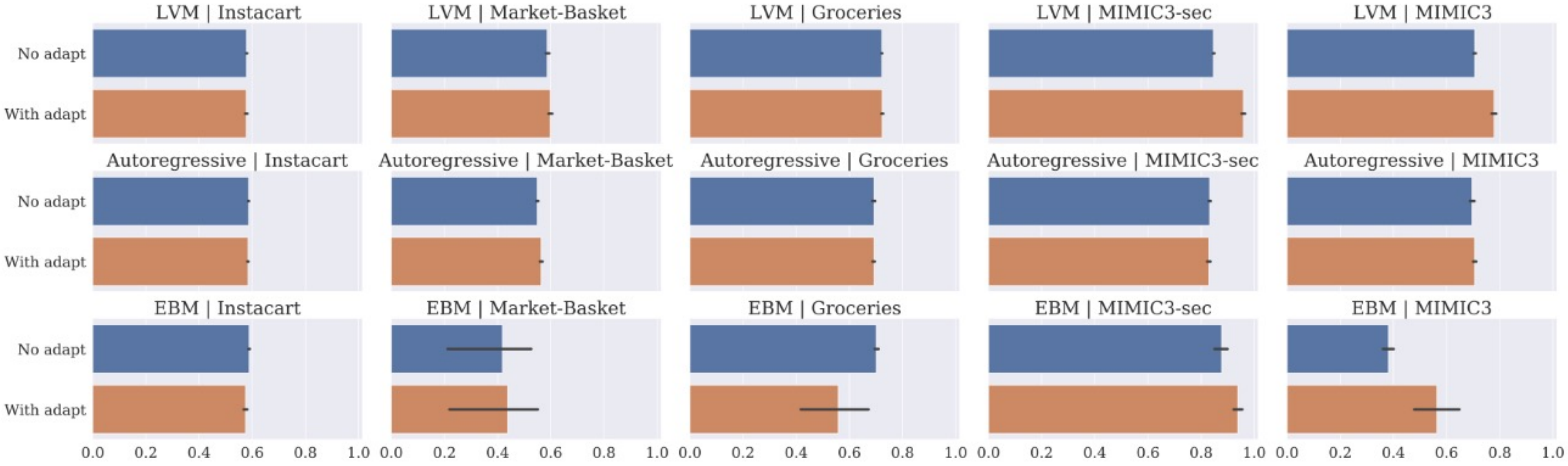


Figure 5. Pairwise-F1 scores for for models before and after marginal adaptations on real-world datasets.

Google Research

Efficiency of adaptation

Table 2. # parameters updated with different methods.

	LVM-continuous	Autoregressive	EBM
(re)training	1,091,239	2,196,657	611,841
MODEM(ours)	512	1,670	167

Table 3. # train/adapt steps until convergence.

(train/adapt)	Groceries	Market-Basket	MIMIC3	MIMIC3-sec	Instacart
LVM	18k/1k	10k/1k	32k/1k	24k/1k	23k/1k
Autoregressive	43k/3k	30k/21k	45k/40k	40k/36k	45k/35k
EBM	99k/14k	60k/10k	62k/12k	95k/5k	105k/12k

Google Research

Thanks

For more information, please feel free to contact us

Contact: hadai@google.com