PAC-Bayes Bounds for Multivariate Linear Regression and Linear Autoencoders

Ruixin Guo [1], Ruoming Jin [1], Xinyu Li [1], Yang Zhou [2]

1. Department of Computer Science, Kent State University
2. Department of Computer Science and Software Engineering, Auburn University

San Diego, December 2-7, 2025

Introduction

In recent years, Linear Autoencoders (LAEs) have demonstrated surprisingly strong, state-of-the-art performance in recommender systems, even outperforming deep neural network models. However,

- The reason behind their strong performance is not well understood.
- Existing works mainly focus on empirical evaluation, offering little theoretical justification.

Statistical learning theory provide a foundation for analyzing model performance.

- Classic uniform convergence PAC bounds are typically too loose for practical use.
- Dziugaite and Roy [1] showed that PAC-Bayes bounds can remain tight even for large models such as deep neural networks, demonstrating their practical value.

Our Goal: Derive PAC-Bayes Bounds to theoretically analyze the performance of LAEs.

Introduction

The Evaluation of LAEs is illustrated as follows:

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Interaction matrix $H \in \{0,1\}^{n\times m}$
with n items and m users.

The Evaluation of LAEs is illustrated as follows:

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Interaction matrix $H \in \{0,1\}^{n \times m}$ with n items and m users.

Split by holding out a fraction of 1s, using a mask matrix $\Delta \in \{0,1\}^{n \times m}$.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & {\color{red}1} \\ 0 & {\color{red}1} & 0 & 0 & 0 & 0 \\ 0 & {\color{red}1} & 0 & 0 & {\color{red}1} & 0 \\ {\color{red}1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & {\color{red}1} & 0 & 0 & 0 \end{bmatrix}$$

Target
$Y = (\mathbf{1} - \Delta) \odot H$
$\in \{0,1\}^{n \times m}$

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & {\color{red}0} \\ 0 & {\color{red}0} & 0 & 0 & 1 & 1 \\ 1 & {\color{red}0} & 0 & 1 & {\color{red}0} & 0 \\ {\color{red}0} & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & {\color{red}0} & 1 & 0 & 1 \end{bmatrix}$$

Input
$X = \Delta \odot H$
$\in \{0,1\}^{n \times m}$

Introduction

The Evaluation of LAEs is illustrated as follows:

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Interaction matrix $H \in \{0,1\}^{n \times m}$ with n items and m users.

Split by holding out a fraction of 1s, using a mask matrix $\Delta \in \{0,1\}^{n \times m}$.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Target
$Y = (\mathbf{1} - \Delta) \odot H \in \{0,1\}^{n \times m}$

$$\begin{bmatrix} 0 & * & * & * & * \\ * & 0 & * & * & * \\ * & * & 0 & * & * \\ * & * & * & 0 & * \\ * & * & * & * & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

LAE model
$W \in \mathbb{R}^{n \times n}$, $\mathrm{diag}(W) = 0$

Input
$X = \Delta \odot H \in \{0,1\}^{n \times m}$

Can be obtained using any training method such as EASE, EDLAE or ELSA.

Introduction

The Evaluation of LAEs is illustrated as follows:

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Interaction matrix $H \in \{0,1\}^{n \times m}$ with n items and m users.

Split by holding out a fraction of 1s, using a mask matrix $\Delta \in \{0,1\}^{n \times m}$.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Target
$Y = (\mathbf{1} - \Delta) \odot H$
$\in \{0,1\}^{n \times m}$

$$\begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix} = \begin{bmatrix} 0 & * & * & * & * \\ * & 0 & * & * & * \\ * & * & 0 & * & * \\ * & * & * & 0 & * \\ * & * & * & * & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Prediction
$WX \in \mathbb{R}^{n \times m}$

LAE model
$W \in \mathbb{R}^{n \times n}$,
$\mathrm{diag}(W) = 0$

Input
$X = \Delta \odot H$
$\in \{0,1\}^{n \times m}$

Can be obtained using any training method such as EASE, EDLAE or ELSA.

The Evaluation of LAEs is illustrated as follows:

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Interaction matrix $H \in \{0,1\}^{n \times m}$ with n items and m users.

Split by holding out a fraction of 1s, using a mask matrix $\Delta \in \{0,1\}^{n \times m}$.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \approx \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & * \\ 0 & * & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & * & 0 \\ * & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & 0 & 0 \end{bmatrix} \Leftarrow \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix} = \begin{bmatrix} 0 & * & * & * & * \\ * & 0 & * & * & * \\ * & * & 0 & * & * \\ * & * & * & 0 & * \\ * & * & * & * & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Target
$Y = (\mathbf{1} - \Delta) \odot H$
$\in \{0,1\}^{n \times m}$

Masked Prediction
$(\mathbf{1} - \Delta) \odot (WX)$

Prediction
$WX \in \mathbb{R}^{n \times m}$

LAE model
$W \in \mathbb{R}^{n \times n}$,
$\mathrm{diag}(W) = 0$

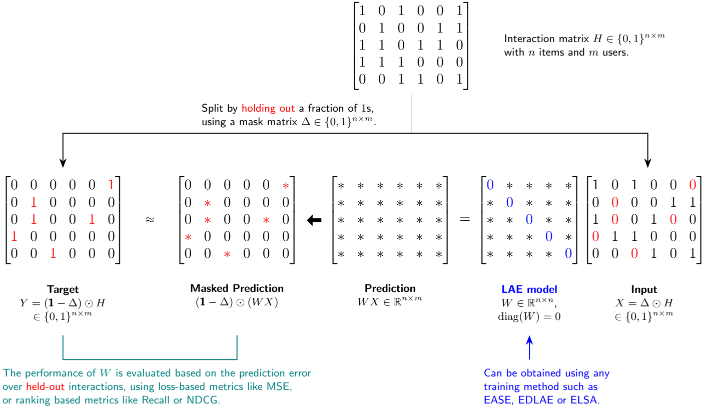
Input
$X = \Delta \odot H$
$\in \{0,1\}^{n \times m}$

The performance of W is evaluated based on the prediction error over held-out interactions, using loss-based metrics like MSE, or ranking based metrics like Recall or NDCG.

Can be obtained using any training method such as EASE, EDLAE or ELSA.

Introduction

The Evaluation of LAEs is illustrated as follows:



Note: When using MSE as the evaluation metric, the error becomes $\|Y - (\mathbf{1} - \Delta) \odot (WX)\|_F^2$, which resembles the multivariate linear regression loss $\|Y - WX\|_F^2$. This highlights the close relationship between multivariate linear regression and LAEs, which motivates our work.

Introduction

Notation:

- Dataset $S = \{(x_i, y_i)\}_{i=1}^m$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}^p$.
- Underlying data distribution: Assume each (x_i, y_i) is i.i.d. drawn from $\mathcal{D}$.
- Input matrix $X = [x_1, ..., x_m]$. Target matrix $Y = [y_1, ..., y_m]$.
- Empirical risk: $R^{\text{emp}}(W) = \frac{1}{m}\|Y - WX\|_F^2 = \frac{1}{m}\sum_{i=1}^m \|y_i - Wx_i\|_F^2$. True risk $R^{\text{true}}(W) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\|y - Wx\|_F^2]$.
- Distribution over W: Prior π. Posterior ρ.

Existing Works:

- Alquier's PAC-Bayes bound [2]: Given π, for any $\lambda > 0$ and $\delta > 0$,

$$P\left(\forall\rho,\, \mathbb{E}_{W\sim\rho}[R^{\text{true}}(W)] < \mathbb{E}_{W\sim\rho}[R^{\text{emp}}(W)] + \frac{1}{\lambda}\left[D(\,\rho\,||\,\pi\,) + \ln\frac{1}{\delta} + \Psi_{\pi,\mathcal{D}}(\lambda, m)\right]\right) \geq 1 - \delta$$

  where $\Psi_{\pi,\mathcal{D}}(\lambda, m) = \ln \mathbb{E}_{W\sim\pi}\mathbb{E}_{S\sim\mathcal{D}^m}[e^{\lambda(R^{\text{true}}(W) - R^{\text{emp}}(W))}]$ and $D(\,\rho\,||\,\pi\,)$ denotes the KL-divergence.

- Shalaeva's bound for single-output linear regression [3], under the Gaussian assumption on $\mathcal{D}$, the term $\Psi_{\pi,\mathcal{D}}(\lambda, m)$ in Alquier's bound can be expressed as:

$$\Psi_{\pi,\mathcal{D}}(\lambda, m) = \ln \mathbb{E}_{W\sim\pi}\frac{\exp(\lambda v_W)}{(1 + \frac{\lambda v_W}{m/2})^{m/2}} \leq \ln \mathbb{E}_{W\sim\pi}\exp\left(\frac{2\lambda^2 v_W^2}{m}\right)$$

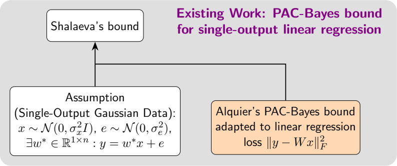
Challenges

To bridge the gap between existing works and our goal, several challenges arise:

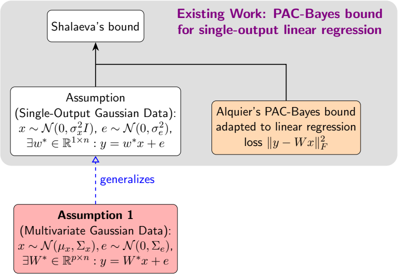
- **Multivariate Data**: Existing PAC-Bayes bounds for linear regression address the single-output case. Extending them to multivariate (multi-output) settings requires more general assumptions that capture dependencies among outputs.
- **LAE-specific Characteristics**: LAEs differ from standard multivariate linear regression in key aspects: bounded data, hold-out constraint between input and target, and zero-diagonal constraint on weight matrix. These characteristics must be formally incorporated into the theoretical analysis.
- **Computational Inefficiency**: Optimizing PAC-Bayes bounds is typically computationally expensive, making it difficult to evaluate them on large models and datasets. Hence, developing more efficient computational methods is critical.

Our work addresses these challenges.

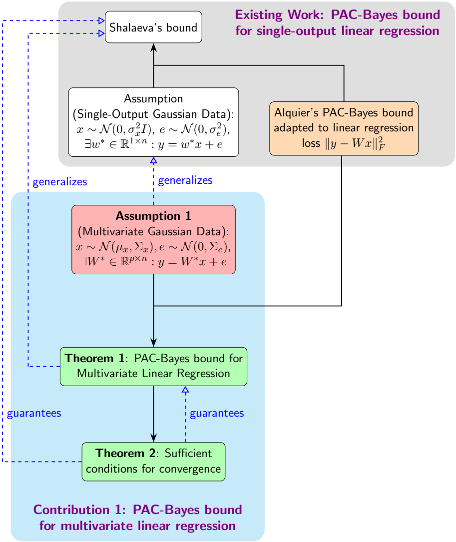
Contributions



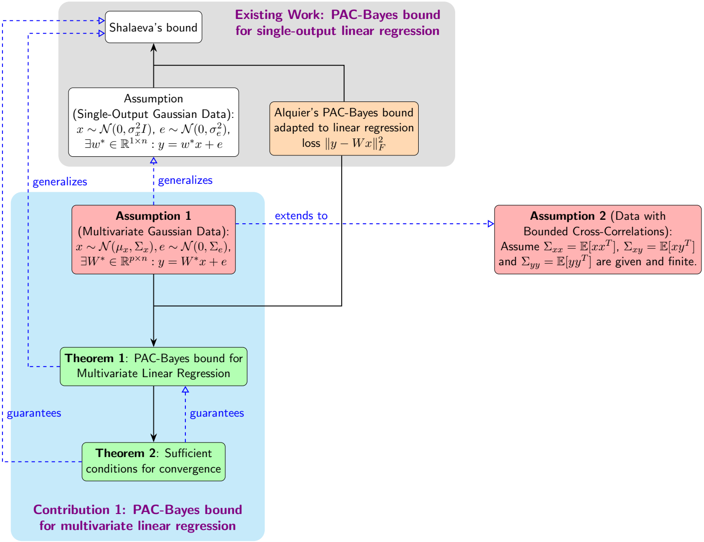
Contributions



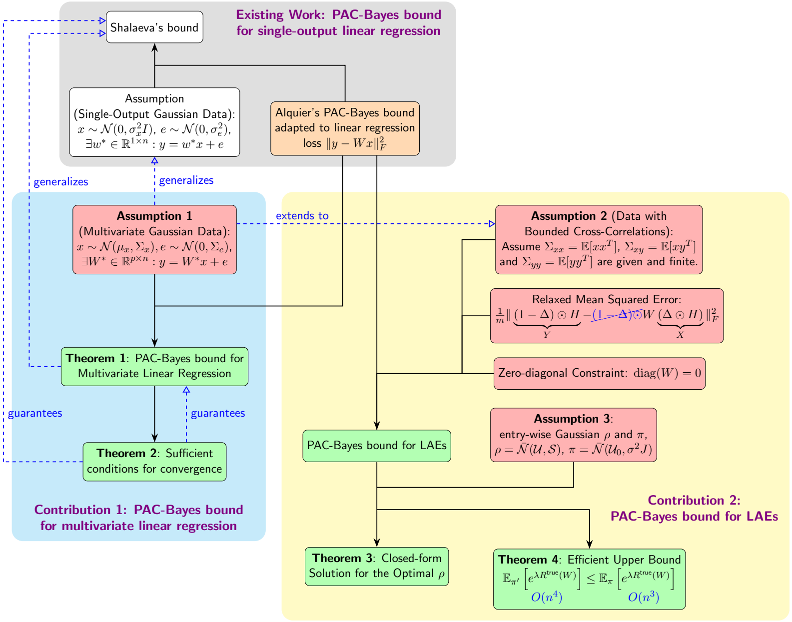
Contributions



Contributions



Contributions



PAC-Bayes Bound for Multivariate Linear Regression

Shalaeva's bound is based on the following assumption.

Assumption: Single-output Gaussian Data

Given $x \in \mathbb{R}^n, y \in \mathbb{R}, e \in \mathbb{R}, \sigma_x, \sigma_e > 0$,
$\mathcal{D}$ satisfies $x \sim \mathcal{N}(0, \sigma_x^2 I)$, $e \sim \mathcal{N}(0, \sigma_e^2)$, $\exists w^* \in \mathbb{R}^{1\times n} : y = w^* x + e$.

Problem: The above assumption cannot be directly adapted to multivariate case.

Solution: We propose a more general assumption first:

Assumption 1: Multivariate Gaussian Data

Given $x \in \mathbb{R}^n, y \in \mathbb{R}^p, e \in \mathbb{R}^p, \mu_x \in \mathbb{R}^n, \Sigma_x \in \mathbb{R}^{n\times n}(PSD), \Sigma_e \in \mathbb{R}^{p\times p}$(PD),
$\mathcal{D}$ satisfies $x \sim \mathcal{N}(\mu_x, \Sigma_x), e \sim \mathcal{N}(0, \Sigma_e)$, $\exists W^* \in \mathbb{R}^{p\times n} : y = W^* x + e$

which reduces to Shalaeva's assumption by taking $p = 1$, $\mu_x = 0$, $\Sigma_x = \sigma_x^2 I$, $\Sigma_e = \sigma_e^2$.

Then drive the PAC-Bayes bound for Multivariate Linear Regression based on it:

Theorem 1

Applying Assumption 1 to Alquier's bound, we get

$$\Psi_{\pi,\mathcal{D}}(\lambda, m) = \ln \mathbb{E}_\pi \left[\exp\left(\lambda\left(\mathrm{tr}(\Sigma_W) + \mu_W^T \mu_W\right)\right) \frac{\exp\left(\sum_{i=1}^p \frac{-\lambda m b_i^2 \eta_i}{m+2\lambda\eta_i}\right)}{\prod_{i=1}^p (1+2\lambda\eta_i/m)^{m/2}} \right] \le \ln \mathbb{E}_\pi \exp\left(\frac{2\lambda^2 \|\Sigma_W\|_F^2}{m}\right)$$

PAC-Bayes Bound for Multivariate Linear Regression

Problem: Shalaeva et al. did not discuss how the choice of π affects convergence, and certain choices of π may fail to guarantee it.

Solution: We propose a sufficient condition that ensures convergence:

Theorem 2

The $\Psi_{\pi,\mathcal{D}}(\lambda, m)$ term converges when (λ, π) satisfy

$$\mathbb{E}_{W\sim\pi}\left[\exp\left(\lambda\|(\Sigma_x + \mu_x\mu_x^T)^{1/2}(W^* - W)\|_F^2\right)\right] < \infty$$

Based on Theorem 2, we can show examples of choices of π that guarantee convergence.

Example

Case 1: If π is of bounded support, then the condition is satisfied for any $\lambda > 0$.

Case 2: If π is entry-wise Gaussian, then there exists $a > 0$ such that for any $\lambda \in (0, a)$, the condition is satisfied.

PAC-Bayes Bound for LAEs

Problem 1: Unlike multivariate linear regression which assumes Gaussian data, LAEs typically operate on bounded data.

Solution: Extend Assumption 1 to account for bounded data.

Assumption 2

Suppose $\mathcal{D}$ is characterized by three finite cross-correlation matrices: $\Sigma_{xx} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[xx^T], \Sigma_{xy} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[xy^T]$ and $\Sigma_{yy} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[yy^T]$.

This assumption holds for all $\mathcal{D}$ with bounded support; it also generalizes Assumption 1.

Problem 2: The classic MSE encodes the hold-out mechanism:

$$\frac{1}{m}\|(\mathbf{1}-\Delta)\odot H - (\mathbf{1}-\Delta)\odot(W(\Delta\odot H))\|_F^2$$

But it differs from the multivariate linear regression loss due to an extra $\mathbf{1}-\Delta$ term.

Solution: Define a relaxed MSE by removing the $\mathbf{1}-\Delta$ term in the classic MSE:

$$\frac{1}{m}\|\underbrace{(\mathbf{1}-\Delta)\odot H}_{Y} - W\underbrace{(\Delta\odot H)}_{X}\|_F^2$$

so that it aligns with the multivariate linear regression loss.

Practical Computation for PAC-Bayes Bound for LAEs

Problem: Not all choices of π and ρ make the optimal bound easy to solve, which can lead to computational inefficiency when evaluating it on large models and datasets.

Solution: Impose the constraint that π, ρ are entry-wise Gaussian:

Assumption 3

Assume π and ρ are entry-wise Gaussian distributions. Given $\mathcal{U}, \mathcal{U}_0, \mathcal{S} \in \mathbb{R}^{n \times n}$ with $\mathcal{S} > 0$ (entry-wise positive) and $\sigma > 0$.

- For $W \sim \pi$, each $W_{ij} \sim \mathcal{N}((\mathcal{U}_0)_{ij}, \sigma^2)$ independently.
- For $W \sim \rho$, each $W_{ij} \sim \mathcal{N}(\mathcal{U}_{ij}, \mathcal{S}_{ij})$ independently.

Under this constraint, the optimal bound is obtained as follows:

Theorem 3

Under Assumption 3, given π, the $(\mathcal{U}, \mathcal{S})$ defining ρ that minimizes the bound admits a closed-form solution.

Compared with the unconstrained case, where the optimal bound is difficult to solve, this constraint allows us to obtain a sub-optimal bound efficiently.

Practical Computation for PAC-Bayes Bound for LAEs

Given that $\Psi_{\pi,\mathcal{D}}(\lambda, m) \leq \ln \mathbb{E}_\pi \left[e^{\lambda R^{\text{true}}(W)}\right]$, computing $\mathbb{E}_\pi \left[e^{\lambda R^{\text{true}}(W)}\right]$ under Assumption 3 costs $O(n^3)$.

Problem: Denote π' as the distribution π with the constraint $\mathrm{diag}(W) = 0$. Then $\mathbb{E}_{\pi'} \left[e^{\lambda R^{\text{true}}(W)}\right]$ has $O(n^4)$ complexity, making it impractical to compute.

Solution: Establish the upper-bound relationship:

Theorem 4

There exists $a > 0$ such that for any $\lambda \in (0, a)$,

$$\mathbb{E}_{\pi'} \left[e^{\lambda R^{\text{true}}(W)}\right] \leq \mathbb{E}_\pi \left[e^{\lambda R^{\text{true}}(W)}\right]$$

And compute $\mathbb{E}_\pi \left[e^{\lambda R^{\text{true}}(W)}\right]$ instead of $\mathbb{E}_{\pi'} \left[e^{\lambda R^{\text{true}}(W)}\right]$, thereby reducing the cost from $O(n^4)$ to $O(n^3)$.

Experimental Results

Experiment Results: Using the EASE LAE Model [4] as an example, we evaluate the gap between the left hand side (LH) and the right hand side (RH) of the bound, as well as the relationship between LH/RH and practical ranking metrics Recall@50/NDCG@100. Results are presented on three datasets: MovieLens 20M, Netflix and MSD.

Models		PAC-Bayes Bound for LAEs			Ranking Performance			
		ML 20M	Netflix	MSD		ML 20M	Netflix	MSD
$\gamma = 50$	LH	61.66	87.22	15.96	Recall@50	0.3434	0.2567	0.3454
	RH	128.66	178.11	32.60	NDCG@100	0.4342	0.3766	0.3187
$\gamma = 100$	LH	60.75	86.54	15.85	Recall@50	0.3453	0.2580	0.3472
	RH	125.90	176.25	32.26	NDCG@100	0.4373	0.3785	0.3205
$\gamma = 200$	LH	60.06	85.96	15.76	Recall@50	0.3471	0.2592	0.3486
	RH	123.67	174.55	31.94	NDCG@100	0.4402	0.3804	0.3220
$\gamma = 500$	LH	59.46	85.35	15.66	Recall@50	0.3489	0.2605	0.3490
	RH	121.41	172.64	31.62	NDCG@100	0.4439	0.3826	0.3225
$\gamma = 1000$	LH	59.19	85.00	15.64	Recall@50	0.3502	0.2612	0.3475
	RH	120.17	171.44	31.50	NDCG@100	0.4464	0.3840	0.3210
$\gamma = 2000$	LH	59.09	84.72	15.68	Recall@50	0.3510	0.2619	0.3434
	RH	119.34	170.45	31.52	NDCG@100	0.4487	0.3854	0.3171
$\gamma = 5000$	LH	59.19	84.48	15.83	Recall@50	0.3506	0.2625	0.3340
	RH	118.91	169.47	31.77	NDCG@100	0.4509	0.3871	0.3079

Conclusion:

- In all cases, RH is within $3\times$ LH, demonstrating that our bound is tight (compared to Dziugaite and Roy [1], where RH $\leq 10\times$ LH).
- Smaller LH/RH correspond to larger Recall/NDCG, indicating the expected correlation and showing that our bound effectively reflects the practical performance of LAE models.

References

[1] Gintare Karolina Dziugaite and Daniel M Roy. *Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data*. UAI, 2017.

[2] Pierre Alquier, James Ridgway, and Nicolas Chopin. *On the properties of variational approximations of Gibbs posteriors*. JMLR, 2016.

[3] Vera Shalaeva, Alireza Fakhrizadeh Esfahani, Pascal Germain, and Mihaly Petreczky. *Improved PAC- Bayesian bounds for linear regression*. AAAI, 2020.

[4] Harald Steck. *Embarrassingly shallow autoencoders for sparse data*. WWW, 2019.

Check **our paper** for more details!
https://openreview.net/pdf?id=S1zkFSby8G

Thank you for attention!