Size- and Dispersion-Corrected Two-Level Softmax Sampling

Walid Bendada and **Guillaume Salha-Galvan**

SPIGM @ ICML 2026 Workshop - walidb@spotify.com

Summary

Context:

- Sampling from a **softmax distribution** is fundamental in ML, but exact sampling scales linearly with the number of items, making it impractical at large scale.
- **Two-level softmax (2LS) sampling** is a popular alternative enabling sublinear-time sampling. Assuming items are partitioned into clusters, 2LS first samples a cluster and then an item within it.

In this paper:

- We show that 2LS introduces systematic **sampling biases** by misweighting clusters, ignoring cluster size imbalance and intra-cluster similarity dispersion.
- We propose **two adjusted sampling methods**: Size-Corrected 2LS (**S-2LS**) and Size- and Dispersion-Corrected 2LS (**SD-2LS**).
- We prove that S-2LS and SD-2LS correct these biases, providing **better softmax approximations** with negligible to non-existent computational overhead.
- **Experiments** on five large-scale datasets validate the improved sampling properties of these methods.

Source code on GitHub:

- `github.com/twolevelsoftmaxanon-creator/2ls`

I - Softmax Sampling

Setup: We consider a set of N items $\mathcal{I} = \{1, \dots, N\}$, each with embedding $x_i \in \mathbb{R}^d$, and a query $q \in \mathbb{R}^d$.

Goal: Sample items according to their similarity to the query q, measured by the dot product $q^\top x_i$.

Softmax Sampling (with temperature $\tau > 0$): popular in ML, but scales linearly w.r.t. N, making it impractical in large-scale settings, e.g., music recommendation at Spotify:

$$\forall i \in \mathcal{I},\ p(i \mid q) = \frac{\exp\left(q^\top x_i/\tau\right)}{\sum_{j=1}^{N} \exp\left(q^\top x_j/\tau\right)} \in [0, 1].$$

II - Two-Level Softmax (2LS) Sampling

Setup: The item set $\mathcal{I}$ is partitioned into $K < N$ disjoint clusters $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$, e.g., using K-means. For each cluster $\mathcal{C}_k$, we define its centroid as $\mu_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} x_i \in \mathbb{R}^d$.

2LS Sampling: first samples a cluster index $k \in \{1, \dots, K\}$, using a softmax over centroids, and then an item $i \in \mathcal{C}_k$ within the selected cluster:

$$p_{\text{2LS}}(k \mid q) = \frac{\exp(q^\top \mu_k/\tau)}{\sum_{k'=1}^{K} \exp(q^\top \mu_{k'}/\tau)} \text{ for } k \in \{1, \dots, K\},$$

$$p_{\text{2LS}}(i \mid k, q) = \frac{\exp(q^\top x_i/\tau)}{\sum_{j \in \mathcal{C}_k} \exp(q^\top x_j/\tau)} \text{ for } i \in \mathcal{C}_k.$$

- Enables sublinear-time sampling, up to $\mathcal{O}(d\sqrt{N})$.
- Factorizes sampling: $p_{\text{2LS}}(i \mid q) = p_{\text{2LS}}(k \mid q)\, p_{\text{2LS}}(i \mid k, q)$.
- Samples over the full set $\mathcal{I}$, in contrast to top-k softmax.
- Closely related to IVF indexing and hierarchical softmax.
- Widely adopted in real-world applications.

III - Size- and Dispersion-Corrected 2LS (S-2LS, SD-2LS) Sampling

S-2LS Sampling: a corrected 2LS procedure that accounts for cluster size imbalance when sampling clusters:

$$p_{\text{S-2LS}}(k \mid q) = \frac{|\mathcal{C}_k| \exp(q^\top \mu_k/\tau)}{\sum_{k'=1}^{K} |\mathcal{C}_{k'}| \exp(q^\top \mu_{k'}/\tau)} \text{ for } k \in \{1, \dots, K\}, \text{then } p_{\text{S-2LS}}(i \mid k, q) = \frac{\exp(q^\top x_i/\tau)}{\sum_{j \in \mathcal{C}_k} \exp(q^\top x_j/\tau)} \text{ for } i \in \mathcal{C}_k.$$

SD-2LS Sampling: further accounts for intra-cluster item-query similarity variance $\sigma_k^2(q) = \mathrm{Var}_{x \sim \mathcal{D}_k}(q^\top x/\tau)$:

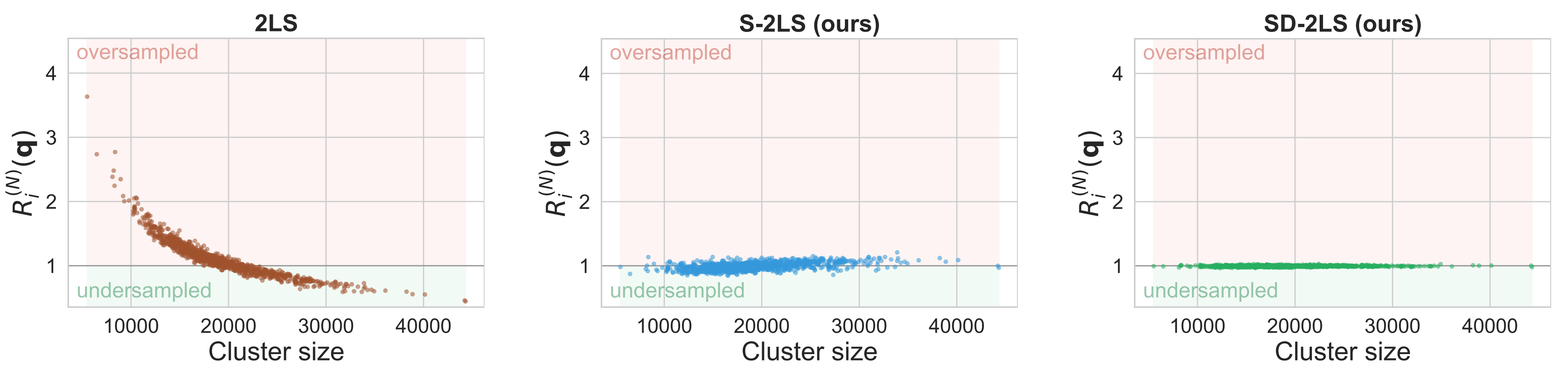
$$p_{\text{SD-2LS}}(k \mid q) = \frac{|\mathcal{C}_k| \exp\left(q^\top \mu_k/\tau + \frac{1}{2}\sigma_k^2(q)\right)}{\sum_{k'=1}^{K} |\mathcal{C}_{k'}| \exp\left(q^\top \mu_{k'}/\tau + \frac{1}{2}\sigma_{k'}^2(q)\right)} \text{ for } k \in \{1, \dots, K\}, \text{then } p_{\text{SD-2LS}}(i \mid k, q) = p_{\text{S-2LS}}(i \mid k, q) \text{ for } i \in \mathcal{C}_k.$$

IV - Theoretical and Experimental Analysis of S-2LS/SD-2LS vs 2LS Sampling

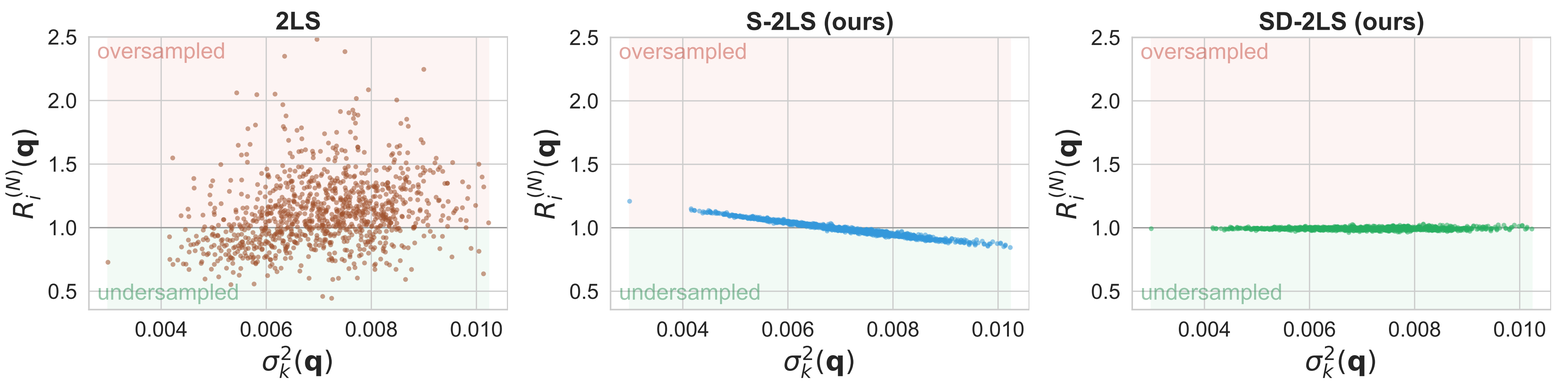
Theoretical Analysis: we introduce the sampling ratio $R_i^{(N)}(q) = \frac{p_{\text{2LS}}(i|q)}{p(i|q)}$ w.r.t. exact softmax and show that:

- 2LS systematically misweights clusters by ignoring cluster size imbalance and intra-cluster similarity dispersion.
- S-2LS corrects the first bias at 2LS complexity. SD-2LS corrects both biases with negligible $\mathcal{O}(d^2\sqrt{N})$ overhead.

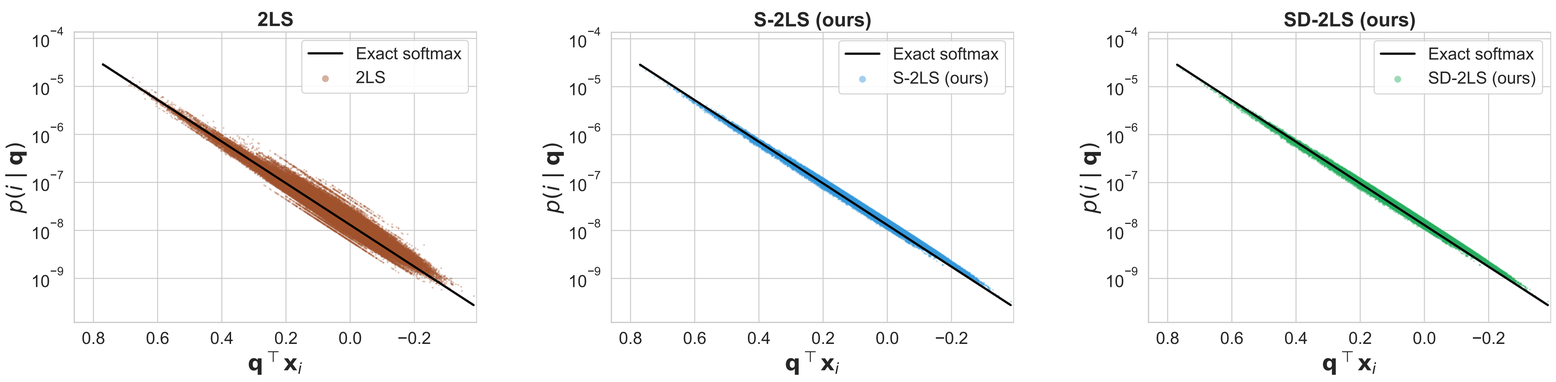
Experimental Analysis: we highlight 2LS biases and S-2LS/SD-2LS corrections on five large-scale datasets:



Sampling ratio on VK-LSVD ($\tau = 0.1$) as a function of cluster size, averaged over queries. Standard 2LS undersamples large clusters and oversamples small ones relative to softmax ($R_i^{(N)}(q) \neq 1$, p-value $p < 0.01$), while S-2LS and SD-2LS correct these biases.



Sampling ratio on VK-LSVD ($\tau = 0.1$) as a function of intra-cluster item-query similarity variance, averaged over queries. The effect is entangled with cluster-size bias for 2LS; S-2LS corrects the cluster-size bias but not the variance bias, isolating the latter and revealing that high- (resp., low-) variance clusters are undersampled (resp., oversampled) relative to softmax ($R_i^{(N)}(q) \neq 1$, p-value $p < 0.01$).



Sampling probability on VK-LSVD ($\tau = 0.1$) per item as a function of query-item similarity for a fixed query. 2LS assigns widely different probabilities to items with identical similarity to the query, while S-2LS and SD-2LS markedly reduce this dispersion.

τ	Dataset	Top-k	2LS	Hierarchical	S-2LS (ours)	SD-2LS (ours)
0.1	GloVe-100	3.9907 ± 0.0395	0.0615 ± 0.0003	0.2897 ± 0.0018	0.0022 ± 0.0002	**0.0001 ± 0.0000**
	VK-LSVD	5.2351 ± 0.0334	0.0297 ± 0.0003	0.3425 ± 0.0027	0.0051 ± 0.0001	**0.0002 ± 0.0000**
	YAMBDA	3.5266 ± 0.0292	0.0300 ± 0.0007	0.6049 ± 0.0154	0.0098 ± 0.0003	**0.0005 ± 0.0000**
	Synth-balanced	0.1580 ± 0.0087	0.3518 ± 0.0436	2.6158 ± 0.0438	0.2334 ± 0.0327	**0.0293 ± 0.0042**
	Synth-unbalanced	0.2318 ± 0.0116	0.2858 ± 0.0391	3.1780 ± 0.0546	0.1728 ± 0.0295	**0.0314 ± 0.0070**
0.2	GloVe-100	5.4103 ± 0.0217	0.0677 ± 0.0001	0.3007 ± 0.0005	0.0001 ± 0.0000	**1.2e-06 ± 9e-08**
	VK-LSVD	7.3333 ± 0.0174	0.0337 ± 0.0001	0.3153 ± 0.0007	0.0003 ± 0.0000	**2.7e-06 ± 1e-07**
	YAMBDA	5.6988 ± 0.0202	0.0388 ± 0.0003	0.3263 ± 0.0020	0.0008 ± 0.0000	**5.3e-06 ± 2e-07**
	Synth-balanced	2.2841 ± 0.0261	0.1296 ± 0.0026	0.6342 ± 0.0029	0.0117 ± 0.0015	**0.0083 ± 0.0021**
	Synth-unbalanced	2.2393 ± 0.0345	0.1935 ± 0.0057	0.9311 ± 0.0070	0.0059 ± 0.0009	**0.0023 ± 0.0011**

Sampling fidelity $\mathrm{KL}(p_{\text{approx}} \,\|\, p_{\text{exact}})$ w.r.t. softmax, averaged over 1000 queries with 95% CIs. *More results on fidelity/latency in the paper.*