The Benefits of Balance: From Information Projections to Variance Reduction

Lang Liu, Ronak Mehta, Soumik Pal, Zaid Harchaoui

UNIVERSITY of WASHINGTON

IFML

NSF

Institute for Foundations of Data Science IFDS Washington · Wisconsin · Santa Cruz · Chicago

Data Balancing

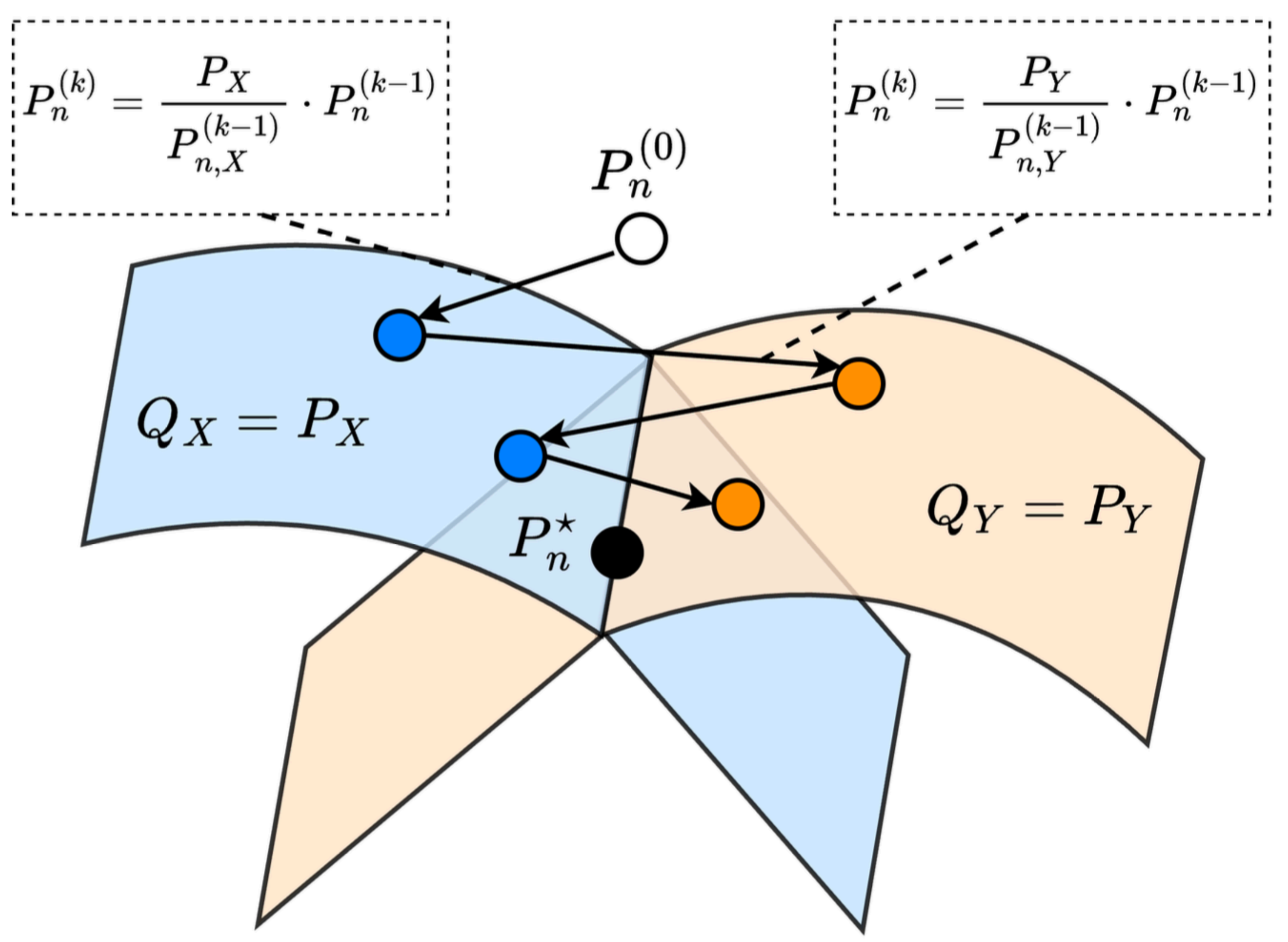
Motivation: High-quality, large-scale datasets of paired observations (features + labels, images + captions) are scarce, while unpaired observations might be abundant.

$$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d}}{\sim} P$$

marginal distributions (**known**) (P_X, P_Y) joint distribution (**unknown**)

How can we incorporate marginal information?

empirical measure $P_n^{(0)} = P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$ marginal likelihood ratio



Estimand	$P(h) = \mathbb{E}_P[h(X,Y)]$
Estimator	$P_n^{(k)}(h) = \mathbb{E}_{P_n^{(k)}}[h(X,Y)]$

How does balancing improve estimation and learning?

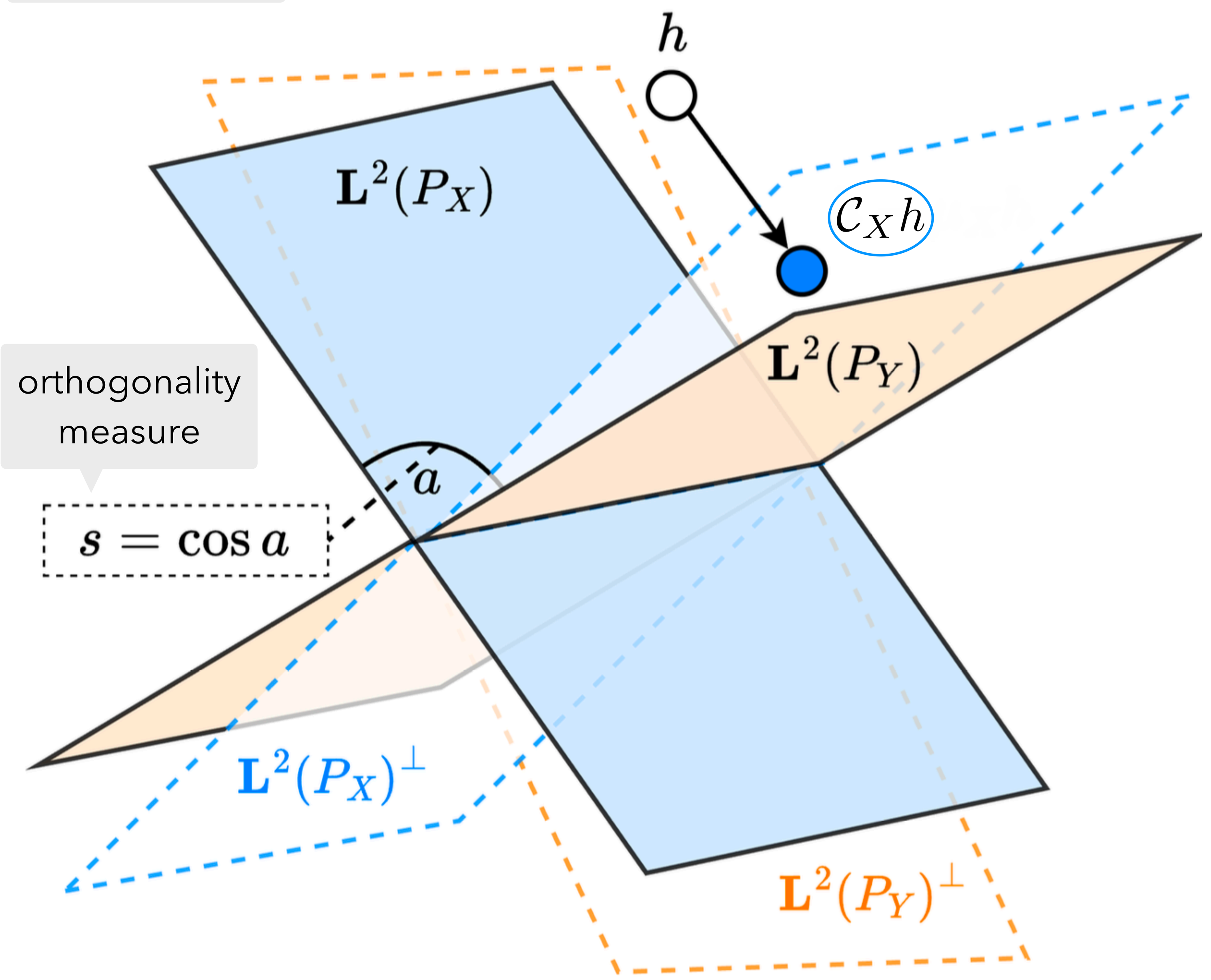
Information Projections → Orthogonal Projections

$$[P_n^{(k)} - P](h) = [P_n^{(k-1)} - P](\mathcal{C}_X h) + O_p(n^{-1})$$
$$= [P_n^{(k-2)} - P](\mathcal{C}_Y \mathcal{C}_X h) + O_p(n^{-1})$$
$$= [P_n - P](\mathcal{C}_Y \mathcal{C}_X \ldots \mathcal{C}_Y \mathcal{C}_X h) + O_p(n^{-1})$$

novel recursion formula

orthogonality measure

$s = \cos a$



Orthogonal Projections → Variance Reduction

We compare the mean squared errors of the empirical versus balanced mean.

$$\sigma^2 = \mathrm{Var}[h(X,Y)] \implies \mathrm{Var}[P_n(h)] = \frac{\sigma^2}{n}$$

Theorem. The iterates of balancing satisfy

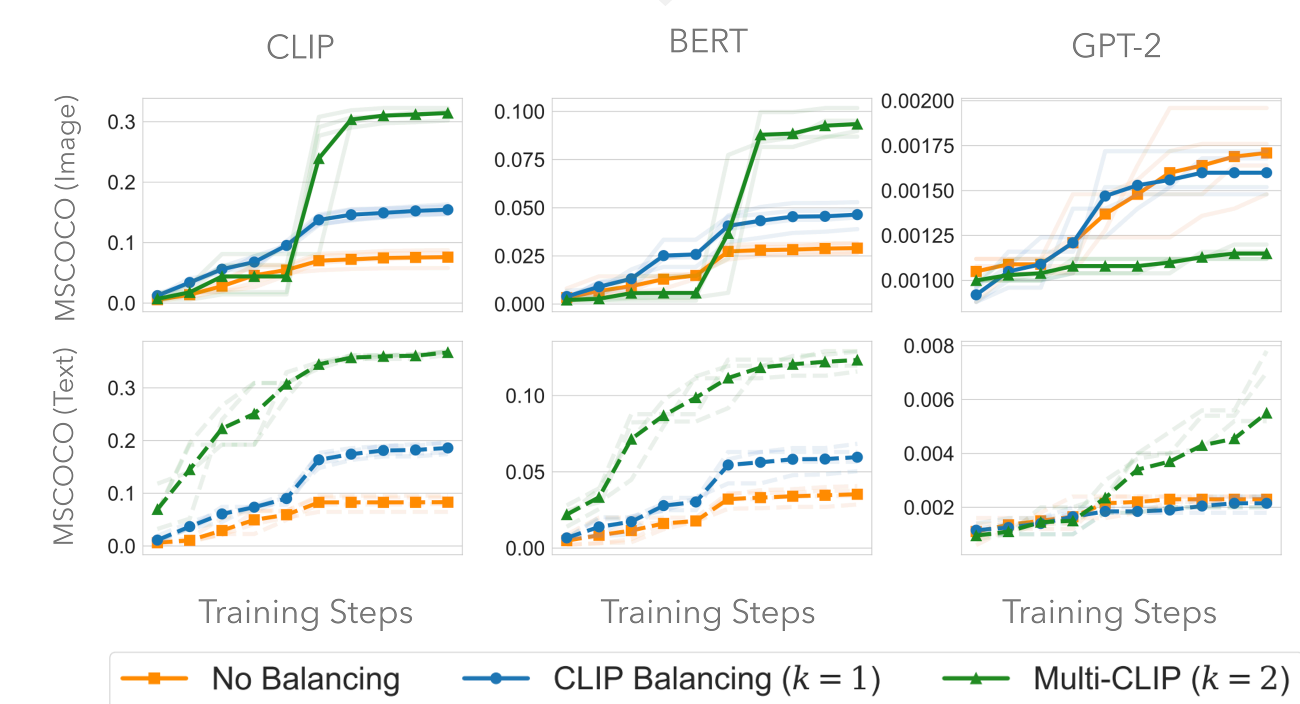
$$\mathbb{E}_P \left|P_n^{(k)}(h) - P(h)\right|^2 = \frac{\sigma^2 - \sigma^2_{\text{gap}}}{n} + O\left(\frac{s^k}{n}\right) + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

The quantity $s \in [0,1)$ can be computed via the spectral properties of the two conditional mean operators.

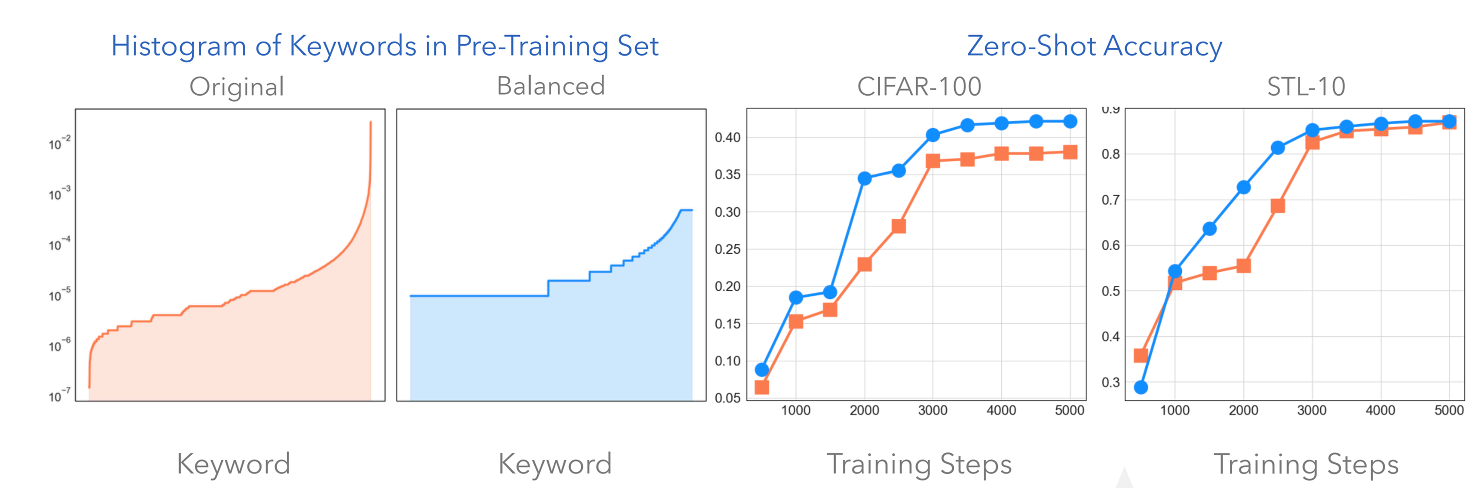
Experiments

Balancing mini-batches to improve the stability of the CLIP training objective.

Using a balanced objective increases zero-shot retrieval (recall) across datasets and embedding architectures.



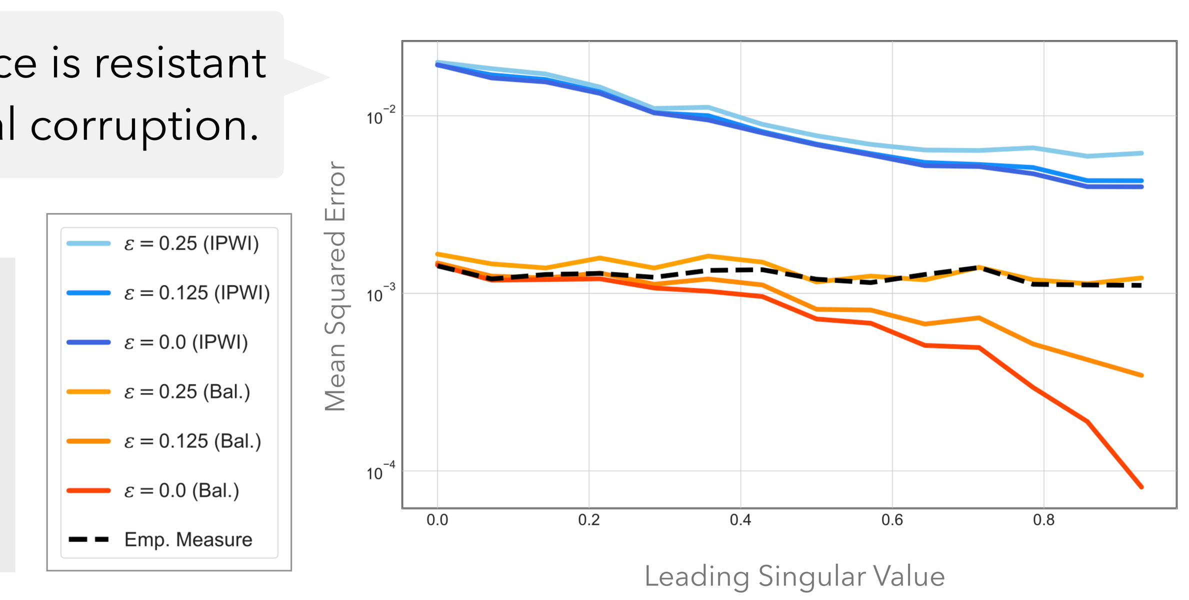
Comparing CLIP models when balancing the entire pre-training set.



Balancing at scale improves performance on zero-shot classification.

Understanding performance under marginal misspecification.

Performance is resistant to marginal corruption.



Code