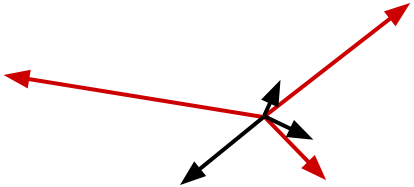
Optimal Batch Variance with Second-Order Marginals

Zelda Mariet, Joshua Robinson, Jamie Smith,
Suvrit Sra, Stefanie Jegelka

Problem setting

Given vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, find

- A distribution p over subsets S of size k
- A mapping μ from S to $\mathbb{R}^d$

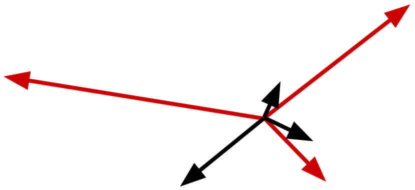


such that

- $\mathbb{E}_{S \sim p}[\mu(S)] = \frac{x_1 + \ldots + x_n}{n}$ (unbiased estimate)
- $\mathbb{E}_{S \sim p}\left[\|\mu(S)\|^2\right]$ is minimized (small variance)

Problem setting

Given vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, find



- A distribution p over subsets S of size k
- A mapping μ from S to $\mathbb{R}^d$

such that

- $\mathbb{E}_{S \sim p}\left[\mu(S)\right] = \dfrac{x_1 + \ldots + x_n}{n}$ (unbiased estimate)
- $\mathbb{E}_{S \sim p}\left[\|\mu(S)\|^2\right]$ is minimized (small variance)

For ex, batched SGD: $\nabla F(x) \approx \frac{1}{k} \sum_{i \in S} \nabla f(x_i)$

$$p = \mathrm{Unif}(1, n, k)$$

$$\mu(S) = \frac{1}{n} \sum_{i \in S} \frac{1}{\Pr(i \in S)} \nabla f(x_i)$$

Theoretical results

✓ The optimal dist. can be characterized by $p_i = \Pr(i \in S),\ p_{ij} = \Pr(\{i,j\} \in S)$

$$p^* \text{ minimizes } \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{p_i}\|x_i\|^2 + \frac{1}{n^2}\sum_{i \neq j}\frac{p_{ij}}{p_i p_j} x_i^\top x_j. \qquad (1)$$

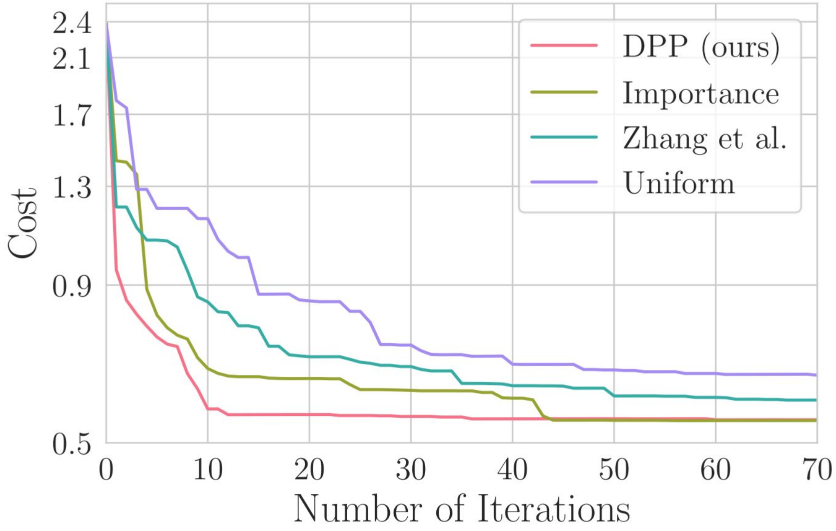
- ✓ Can be used to compare different distributions
- ✓ Can be used to learn parametric distributions (*e.g.*, DPPs)

✓ For sequential independent samples with replacement, importance sampling is optimal: $p_i \propto 1/\|x_i\|$

x Reverse engineering (1) is NP-hard

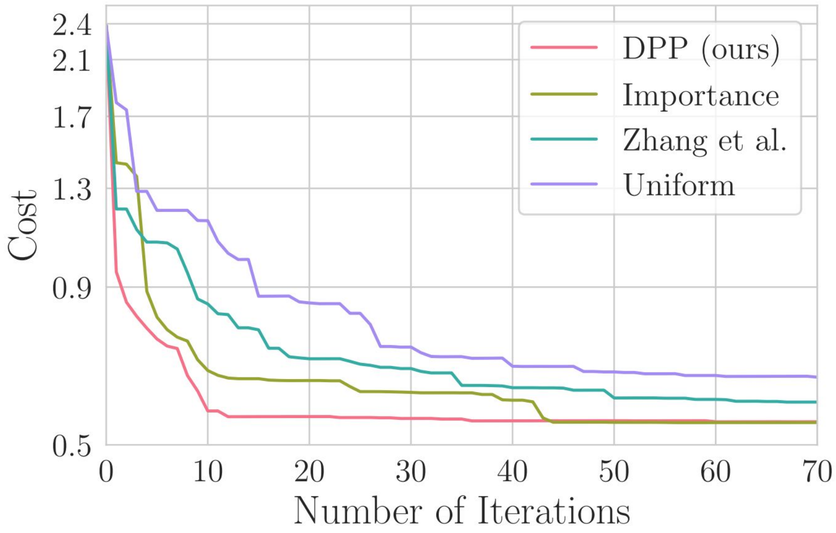
Experimental results

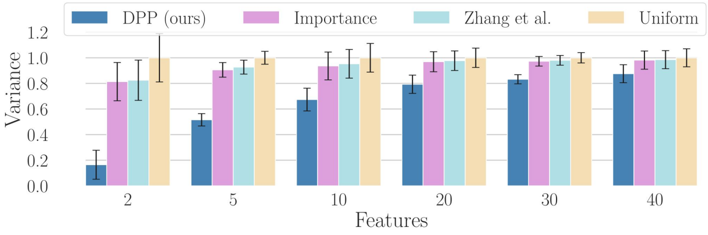
Learning a DPP over gradient batches without replacement:



Experimental results

Learning a DPP over gradient batches without replacement:





Future work:

- Leverage approximate bounds for $x_i^\top x_j$
  [Zhao & Zhang, ICML'15; Loshchilov & Hutter, '15; Katharopoulos & Fleuret, ICML'18]
- Application to active learning