ETH zürich

Working Title

Max Paulus

Introduction I

Given a set input $X = \{X_i\}_{i=1}^{n}$ and an associated target $Y \in \mathbb{R}^d$, jointly learn a determinantal point process $\mathbb{P}_L(\mathcal{A})$, where $\mathcal{A}$ indexes subsets of X and a function $g : 2^X \to \mathbb{R}^d$, such that $\hat{Y} = g(\mathcal{A})$ and $\mathrm{cost} = \mathrm{cost}(Y, \hat{Y})$ is minimized.

Architecture

- Kernel Network: $L = VV^\top$, where $V_{i,\cdot} = f_\theta(X_i, \overline{X})$
- Prediction Network: $\hat{Y} = g(\overline{X}_\mathcal{A})$, where $\mathcal{A} \subseteq X$
- Sampling: Use SVD of V to compute eigenvalues and -vectors of L.

Gradient Derivation

Let θ parameterise the kernel network, such that $K = f_\theta(X)$ and let $L = KK^\top$ parameterise the DPP. The policy gradient is given by:

$$\nabla_\theta \mathbb{E}_L[\mathrm{cost}] = \mathbb{E}_L[\nabla_\theta \log(\mathbb{P}(\mathcal{A})) \times \mathrm{cost}] \tag{1}$$

We need $\nabla_V \log(\mathbb{P}(\mathcal{A}))$ [1]:

$$\begin{aligned} \nabla_V \log(\mathbb{P}(\mathcal{A})) &= \nabla_V \log\det(L_\mathcal{A}) - \nabla_V \log\det(L + I) \\ &= 2 \times L_\mathcal{A}^{-1} K - 2 \times (I_n - K(I_d - +K^\top K)^{-1} K^\top) \end{aligned}$$

Why good?

- The dimension of $L_\mathcal{A}$ and $(I_d - +K^\top K)^{-1})$ do not depend on ground set size, can be computed even for large sets.

Controlling the Variance

VIMCO [3] is a **state-of-the-art leave-one-out control variate** for multi-sample MC objectives. Can also be (ab)used for additive decomposable loss function and provide high-quality baseline:

$$\mathrm{cost}(\mathcal{A}_i) \rightarrow \mathrm{cost}(\mathcal{A}_i) - \frac{1}{n-1}\sum_{-i}\mathrm{cost}(\mathcal{A}_i) \tag{2}$$

Why good?

- Unbiased
- No extra parameters
- Credit assignment (preserved)
- Loss scaling

Controlling Sparsity

Given $L = KK^\top$ and $K = USV^\top$, $\mathbb{E}[|\mathcal{A}|]$ of a sampled subset $\mathcal{A}$ is [2]:

$$\mathbb{E}[|\mathcal{A}|] = \mathrm{Tr}(L(L+I)^{-1})$$
$$= \sum_{i=1}^{n} \frac{S_i^2}{S_i^2 + 1}$$

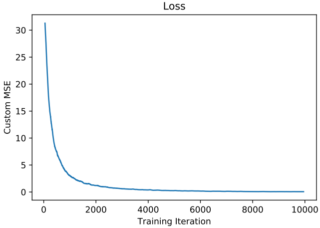
Why bother?

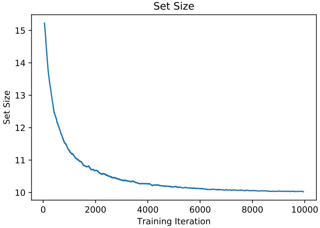
- **Naive:** Regularise subset size of samples directly through REINFORCE
- **Using above:** Expectation is tractable; backpropagation through singular values, reduces variance and increases quality of learning signal for policy gradient.

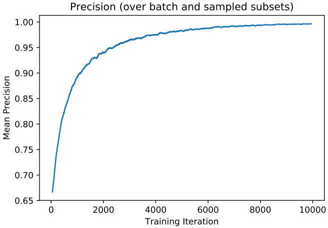
Learning a k-DPP - Set-Up

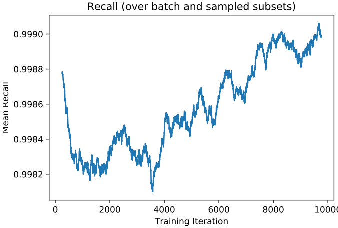
- **Task:** Given sets of size 40 with each member drawn from one of 10 clusters, learn a 10-DPP that always selects one and only one member from each cluster. Cluster means $\in \mathcal{Z}^{50}_{[-50,50]}$
- **Loss:** Use direct supervision on returned subset and a high-quality learning signal: $(\#\text{missed} + \#\text{oversampled})^2$
- **Network:** Uses only a 2-hidden layer kernel network with dimensions [100, 500, 500, 100]
- **Training:** Iterations: 10k, Batchsize: 10, Learning rate: 1^{-5}, Samples: 4, Optimizer: ADAM

Learning a k-DPP - Results I









Learning a k-DPP - Results II

	learnt DPP	random benchmark
Loss	0.03	52.46
Clusters missed	0.08%	34.42%
Clusters oversampled	0.21%	26.60%
Mean(Subset Size)	10.01	10.06
Var(Subset Size)	0.03	7.52
Perfect Cluster returned	97.2%	0.0%

Outlook

Architecture

- Separate quality and diversity models in kernel network
- Explore successful application of Deep Set architecture

Training

- Alternative sampling distribution to increase exploration (marginals?)
- Could explore loss-scale invariant signal through suitable transformation
- Demonstrate superiority of control variate and regularization

Applications

- Multi-Sentiment Prediction?
- Similar Question Retrieval?
- Recommender Systems?
- ...

M. Gartrell, U. Paquet, and N. Koenigstein.
Low-Rank Factorization of Determinantal Point Processes for Recommendation.
ArXiv e-prints, February 2016.

Alex Kulesza.
Learning with Determinantal Point Processes.
PhD thesis, University of Pennsylvania, 2012.

Andriy Mnih and Danilo Jimenez Rezende.
Variational inference for monte carlo objectives.
CoRR, abs/1602.06725, 2016.