# Supplementary Material

## Selecting points from a plane with DPP

To illustrate how DPP works as a weighting scheme for a greedy selection algorithm we use an artificial task of selecting points from a plane. The source code for the problem is available online at `https://gist.github.com/eiennohito/a669b296ac4cb1be9542ef53f0810b6d`. You need to have a Python 2.7 with numpy and matplotlib packages installed to run it.

We use a grid of $20 \times 20$ evenly spaced points for this task. Similarity of two points $p_i$ and $p_j$ is be computed from their distance as

$$\mathrm{sim}(p_i, p_j) = \frac{d_m - \mathrm{dist}(p_i, p_j)}{d_m},$$

where $d_m = \max \mathrm{dist}(p_i, p_j)$ for any pair of $p_i$ and $p_j$. Distance between two points is a Euclidean distance between them. The distance is used as elements of $L$-kernel:

$$L_{i,j} = \mathrm{sim}(p_i, p_j).$$

Remember that the feature representation in the main paper uses elements of $L$-kernel as $L_{ij} = q_i \phi_i^T \phi_j q_j$ where $q_i$ are scalars and $\phi_i$ are unit-length vectors. We use the similarity based on euclidian distance instead of cosine similarity $\phi_i^T \phi_j$ for our simple demonstration.

Normalized item selection marginal probability heatmap is shown on Figure 1. Top-left image is for initial probability estimation. Other images are created after some points are selected. It is possible to observe that DPP assigns higher probabilities to items on borders and especially in corners.

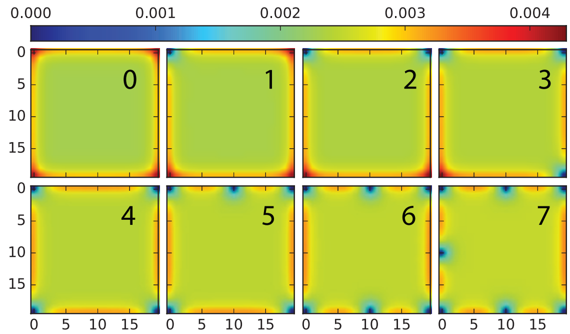


Figure 1: Item Selection Marginal Probabilities Heatmap: diversity features only. On each step an item with highest probability is selected. Places with zero probability are from selected items.

Reason for that is simple. Items in corners have highest distance from each other, which means that similarity for them is going to be the lowest. Probability of selecting a set of items according to DPP is proportional to determinant of indexed $L$ kernel. The determinant value is going to be higher if the off-diagonal elements are smaller, which is exactly the case for points in the corners of the rectangle.

Another observation is that regions of lower probability near selected items are relatively small. Authors of the DPP paper add a parameter $r$ that forces every item to be more similar for their tasks. Intuitively, each point here *is* a point and should not be very different from other points. Using this idea, the $L$-kernel items are going to be computed as

$$L_{i,j} = r + (1 - r)\,\mathrm{sim}(p_i, p_j)$$

for the artificial toy task.

For the value of $r = 0.9$ selection heatmaps are shown on the Figure 2. In this setting it is possi-

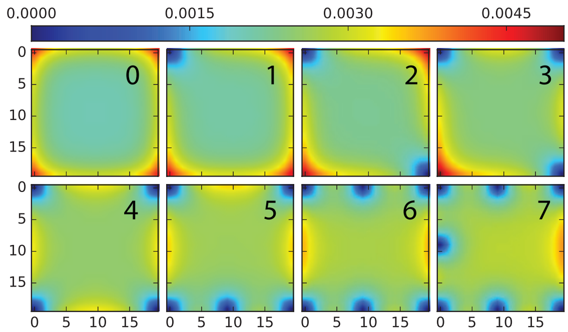


Figure 2: Selection Marginal Probabilities Heatmap: diversity features, $r = 0.9$

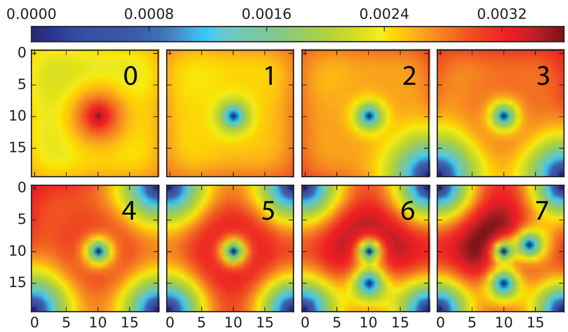


Figure 3: Selection Marginal Probabilities Heatmap: diversity and quality features, $r = 0.9$

ble to see that compared to the Figure 1, changes in probabilities after each selection became much rapid. Basically the larger values of parameter $r$ makes DPP to discount larger vicinities of a selected item when considering the next selection.

Still, it is possible to say that the greedy algorithm on diversity features only is going to select non-similar items, however they are not going to be *representative*. Furthermore, selected items are going to be outliers, which is not a useful property for selecting example sentences. For the task of summarization, authors of the DPP paper use features that serve a measure of centrality as *quality* features.

Let us see how quality features can change probability distribution. For each item let's introduce a scalar value $q_i$ that is going to be close to 1 if the item is closer to the center of the plane and decrease with the distance from the center of the plane. One possibility is to have

$$q_i = \exp(-\operatorname{dist}(p_i, p_0)),$$

where $p_0$ is the center of the plane. Values of $L$-kernel are going to be

$$L_{i,j} = r + (1 - r)\operatorname{sim}(p_i, p_j) q_i q_j.$$

The Figure 3 shows point selection marginal probability heatmap for the case of using both types of features and the parameter $r = 0.9$. This time the selection starts from the center of the plane which is more useful for selecting items which are non-similar and representative at the same time. For the cases when DPP is used to select items which are central and non-similar at the same time, usage of a centrality-like measure in quality features is essential.