F.1 Datasets

We considered these five datasets:

A A least squares problem with a data matrix $A \in \mathbb{R}^{m \times n}$ and target $b \in \mathbb{R}^m$,

$$\arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2.$$

We set A to be an m by n matrix with entries sampled from a $\mathcal{N}(0,1)$ distribution (with $m = 1000$ and $n = 10000$). We then added 1 to each entry (to induce a dependency between columns), multiplied each column by a sample from $\mathcal{N}(0,1)$ multiplied by ten (to induce different Lipschitz constants across the coordinates), and only kept each entry of A non-zero with probability $10\log(m)/m$. We set $b = Ax + e$, where the entries of x and e were drawn from a $\mathcal{N}(0,1)$ distribution.

B A binary logistic regression problem of the form

$$\arg\min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} \log(1 + \exp(-b_i x^T a_i)).$$

We use the data matrix A from the previous dataset (setting row i of A to a_i^T), and b_i to be the sign of $x^T a_i$ using the x used in the generating the previous dataset. We then flip the sign of each entry in b with probability 0.1 to make the dataset non-separable.

C A multi-class logistic regression problem of the form

$$\arg\min_{x \in \mathbb{R}^{d \times k}} \sum_{i=1}^{m} \left[-x_{b_i}^T a_i + \log\left(\sum_{c=1}^{k} \exp(x_c^T a_i) \right) \right],$$

see (38). We generate A as in the previous two cases. To generate the $b_i \in \{1, 2, \ldots, k\}$ (with $k = 50$), we compute $AX + E$ where the elements of the matrices $X \in \mathbb{R}^{d \times k}$ and $E \in \mathbb{R}^{m \times k}$ are sampled from a standard normal distribution. We then compute the maximum index in each row of that matrix as the class labels.

D A label propagation problem of the form

$$\min_{x_i \in S'} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - x_j)^2,$$

where x is our label vector, S is the set of labels that we do know (these x_i are in $\{-1, 1\}$), S' is the set of labels that we do not know, and $w_{ij} \geq 0$ are the weights assigned to each x_i describing how strongly we want the labels x_i and x_j to be similar. We set the non-zero pattern of the w_{ij} so that the graph forms a 50 by 50 lattice-structure (setting the non-zero values to 10000). We labeled 200 points, leading to a problem with 2300 variables but where each variable has at most 4 neighbours in the graph.

E Another label propagation problem for semi-supervised learning in the 'two moons' dataset [99]. We generate 500 samples from this dataset, randomly label five points in the data, and connect each node to its five nearest neighbours (using $w_{ij} = 1$). This results in a very sparse but unstructured graph.

F.2 Greedy Rules with Gradients Updates

In Figure F.2 we show the performance of the different methods from Section 7.1 on all five datasets with three different block sizes. In Figure F.2 we repeat the experiment but focusing only on the FB methods.