Statistical Data Science - Home assignment 3

Prof. Dr. Philipp Otto

> **Note**
>
> issue date :
> Submission date :
> Name, matriculation number:
> Evaluation:

Problem - Proof decision problems

Proof that almost every decision problem is not solvable in finite time!

Problem - Time complexity

Compute the asymptotic (time) complexity of the following code fragments:

a)

```
sum = 0;
for(int i = 0; i < n; i++)
{
    sum += i;
}
```

b)

```
sum = 0;
for(int i = 0; i < n; i++)
{
    sum += i;
    for(int j = 0; j < n; j++)
    {
        sum += j;
    }
}
```

c)

```
void matrix_multiplication(double MatrixA[p][p], double MatrixB[p][p],
    double MatrixAB_product[p][p])
{
  for (int i = 0; i < p; i++)
  {
    for (int j = 0; j < p; j++)
    {
      double sum_k = 0;
      for (int k = 0; k < p; k++)
      {
        sum_k = sum_k + MatrixA[i][k] * MatrixB[k][j];
      }
      MatrixAB_product[i][j] = sum_k;
    }
  }
}
```

Problem - P versus NP

Explain the P versus NP problem!

Problem - Maximum speedup

Let p be the number of parallel processes and π denotes the fraction of code that can be parallelized.

- Compute an upper bound for the speed-up that can be obtained by parallelizing the code.
- Visualize the (theoretical) speed-up for different values of p and π.
- Explain the difference between weak and strong scaling.

Furthermore, perform a Monte Carlo simulation study to illustrate the speed-up, which can achieved by parallelizing. For this reason, simulate 10^8 gaussian random numbers with $\mu = 1$ and $\sigma = 2$. Distribute the simulation on $p = 1$ and $p = 2$ parallel processes. Estimate the required computation time with $m = 1000$ replications.

Problem - Model selection

In Figure , the efficiency of several models is given for the empirical example considered in Vetter, P., Schmid, W., Schwarze, R. (2016). Which model is the best under the following restrictions:

1. $\Sigma^{-1}\tilde{Z}$ should be computed in less than 0.01 seconds
2. $\Sigma^{-1}\tilde{Z}$ should be computed in less than 0.2 seconds
3. $\Sigma^{-1}\tilde{Z}$ should be computed in less than 0.5 seconds
4. the mean squared prediction error should be less than 2 and $\Sigma^{-1}\tilde{Z}$ should be computed as fast as possible

Table 3: Efficiency Evaluation - Subset

Model Type	Parameters r	γ	MSPE	Time in sec $\mathbf{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	Time in sec $\det \mathbf{\Sigma}$	Memory in GB $\mathbf{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	Memory in GB $\det \mathbf{\Sigma}$
Full Model			0.975	11.0866	12.7267	2.4645	2.3657
Fixed	10	0	3.194	0.00111	0.00037	0.00099	0.00033
Rank	42	0	2.684	0.01967	0.00656	0.01745	0.00574
Kriging	132	0	1.795	0.18632	0.07271	0.17236	0.05666
	0	50	3.216	0.00001	0.00001	0.00001	0.00001
	0	100	3.186	0.00005	0.00003	0.00005	0.00002
	0	200	3.055	0.00067	0.00041	0.00072	0.00024
	0	300	2.783	0.00304	0.00185	0.00325	0.00107
Covariance	0	400	2.466	0.00826	0.00526	0.00900	0.00296
Tapering	0	500	2.192	0.01758	0.01197	0.01966	0.00646
	0	625	1.963	0.04159	0.02916	0.04708	0.01548
	0	750	1.734	0.06561	0.04635	0.07450	0.02449
	0	1000	1.482	0.15570	0.11324	0.17895	0.05883
	0	1500	1.241	0.47393	0.35510	0.55164	0.18134
	10	50	3.184	0.00113	0.00038	0.00100	0.00033
	10	100	3.155	0.00115	0.00041	0.00104	0.00034
	10	200	3.026	0.00182	0.00074	0.00171	0.00056
	10	300	2.759	0.00438	0.00200	0.00424	0.00140
Full-scale	10	400	2.447	0.01292	0.00209	0.00999	0.00328
Approximation	10	500	2.177	0.02048	0.01055	0.02065	0.00679
(r=10)	10	625	1.951	0.04693	0.02531	0.04807	0.01580
	10	750	1.726	0.07337	0.04008	0.07549	0.02481
	10	1000	1.477	0.16990	0.10052	0.17994	0.05915
	10	1500	1.239	0.51520	0.31532	0.55263	0.18167
	42	50	2.676	0.01968	0.00656	0.01746	0.00574
	42	100	2.654	0.01958	0.00673	0.01750	0.00575
	42	200	2.558	0.02024	0.00706	0.01817	0.00597
	42	300	2.358	0.02266	0.00845	0.02070	0.00681
Full-scale	42	400	2.121	0.02801	0.01173	0.02645	0.00869
Approximation	42	500	1.912	0.03929	0.01648	0.03711	0.01220
(r=42)	42	625	1.734	0.06663	0.03035	0.06453	0.02121
	42	750	1.556	0.09396	0.04422	0.09195	0.03023
	42	1000	1.356	0.19425	0.10091	0.19640	0.06456
	42	1500	1.163	0.55367	0.30158	0.56909	0.18708
	132	50	1.791	0.18665	0.07240	0.17237	0.05666
	132	100	1.780	0.18651	0.07260	0.17241	0.05668
	132	200	1.736	0.18797	0.07214	0.17308	0.05690
	132	300	1.644	0.19186	0.07206	0.17561	0.05773
Full-scale	132	400	1.531	0.19894	0.07361	0.18136	0.05962
Approximation	132	500	1.429	0.21101	0.07757	0.19202	0.06312
(r=132)	132	625	1.338	0.23900	0.09079	0.21944	0.07214
	132	750	1.247	0.26698	0.10401	0.24686	0.08115
	132	1000	1.140	0.37446	0.15351	0.35131	0.11549
	132	1500	1.034	0.76104	0.32702	0.72400	0.23800

Figure 1: Vetter, P., Schmid, W., Schwarze, R. (2016). Efficient approximation of the spatial covariance function for large datasets - analysis of atmospheric CO2 concentrations, Journal of Environmental Statistics, 6(3); Table 3