Variance Reduction for Inverse Trace Estimation via Random Spanning Forests

Yusuf Yiğit PİLAVCI, Pierre-Olivier AMBLARD, Simon BARTHELMÉ, Nicolas TREMBLAY

Introduction

- Trace (sum of the diagonal entries of a matrix) is an essential algebraic operation.
- In many applications, $\mathrm{tr}(f(\mathsf{L}))$ is the quantity of interest for a given matrix L.
- In this work, we focus on

$$f(\mathsf{L}) = q(\mathsf{L}+q\mathsf{I})^{-1},$$

where $q>0$ and L is symmetric and diagonally dominant.

Hyperparameter Selection

Regularized Regression on Graphs

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{z}\in\mathbb{R}^n} \underbrace{q\,||\mathbf{y}-\mathbf{z}||^2}_{\text{Fidelity}} + \underbrace{\mathbf{z}^T\mathsf{L}\mathbf{z}}_{\text{Regularization}}, \quad q>0$$

where $\mathbf{y}\in\mathbb{R}^n$ is a graph signal. L denotes the graph Laplacian of the given graph and q is the regularization parameter.

The explicit solution to this problem is:

$$\hat{\mathbf{x}} = \mathsf{K}\mathbf{y} \text{ with } \mathsf{K} = q(\mathsf{L}+q\mathsf{I})^{-1}$$

where I is the identity matrix.

- The denoising error $||\mathbf{x}-\hat{\mathbf{x}}||_2^2$ highly depends on q.
- Many methods of finding good value of q needs to compute $\mathrm{tr}(\mathsf{K})$. For example, generalized cross-validation computes:

$$GCV(q) = \frac{1}{N}\left(\sum_{i=1}^{n} \frac{y_i-\hat{x}_i}{1-(\mathrm{tr}(\mathsf{K})/n)}\right).$$

- However, computing the inverse takes $\mathcal{O}(n^3)$ operations.

State-of-the-Art

Hutchinson's estimator

The state-of-the-art algorithm for estimating $\mathrm{tr}(\mathsf{K})$ is:

$$h := \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}^{(i)\top}\mathsf{K}\mathbf{a}^{(i)}$$

where $\mathbf{a}^{(i)}\in\{-1,1\}^n$ is a random vector with $\mathbb{P}(\mathbf{a}_j^{(i)}=\pm 1)=1/2$.

- h is unbiased for estimating $\mathrm{tr}(\mathsf{K})$.
- The cumbersome computation is $\mathsf{K}\mathbf{a}^{(i)}$ for N vectors.
- It can be done via sparse Cholesky decompisition, iterative solvers, algebraic multigrid solvers, fast solvers for Laplacian systems...

Random Spanning Forest based Estimators

For an undirected graph $\mathcal{G}=(\mathcal{V},\mathcal{E},\mathsf{W})$:

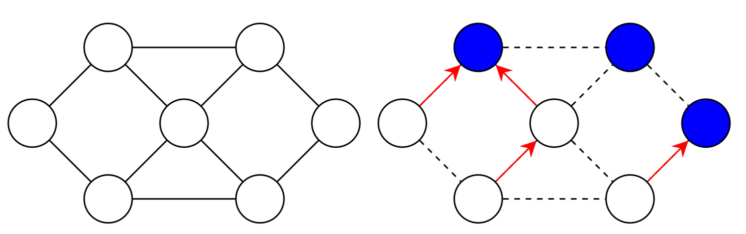


Fig. 1: Original graph and a rooted spanning forest

Random Spanning Forests (RSF)

Consider the following parametric distribution over rooted spanning forests:

$$P(\Phi_q=\phi) \propto q^{|\rho(\phi)|}\prod_{\tau\in\phi}\prod_{(i,j)\in\tau} \mathsf{W}_{i,j}$$

where q is a parameter and $\rho(\phi)$ denotes the set of roots in the forest ϕ. One can sample from this distribution by a variant of Wilson's algorithm in time $\mathcal{O}(|\mathcal{E}|/q)$.

A key result:

$$\mathbb{E}[|\rho(\Phi_q)|] = \mathrm{tr}(\mathsf{K}) \text{ with } \mathrm{Var}(|\rho(\Phi_q)|) = \mathrm{tr}(\mathsf{K}-\mathsf{K}^2).$$

Previously, $|\rho(\Phi_q)|$ is used for estimating $\mathrm{tr}(\mathsf{K})$. In this work, we improve its performance.

Proposed Methods

Control Variate Method

Estimation of K can be considered as minimizing the loss following function:

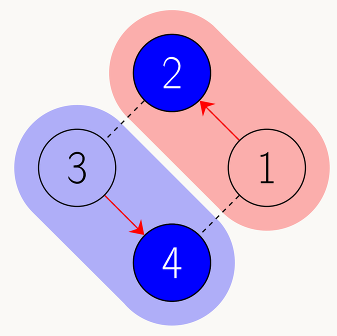
$$L(\mathsf{S}) = \mathrm{tr}\left(\frac{1}{2}\mathsf{S}^\top\mathsf{K}^{-1}\mathsf{S}-\mathsf{S}\right)$$

The gradient descent algorithm draws the following iteration:

$$\mathsf{S}_{k+1} = \mathsf{S}_k - \alpha(\mathsf{K}^{-1}\mathsf{S}_k-\mathsf{I}).$$

where α is the update size.

In our previous work, we give two unbiased estimators $\tilde{\mathsf{S}}$ and $\bar{\mathsf{S}}$:



$$\tilde{\mathsf{S}} = \begin{array}{c} \\ 1\\2\\3\\4\end{array}\!\!\begin{array}{c}\begin{matrix}1&2&3&4\end{matrix}\\ \begin{bmatrix}0&1&0&0\\0&1&0&0\\0&0&0&1\\0&0&0&1\end{bmatrix}\end{array} \qquad \bar{\mathsf{S}} = \begin{array}{c} \\ 1\\2\\3\\4\end{array}\!\!\begin{array}{c}\begin{matrix}1&2&3&4\end{matrix}\\ \begin{bmatrix}1/2&1/2&0&0\\1/2&1/2&0&0\\0&0&1/2&1/2\\0&0&1/2&1/2\end{bmatrix}\end{array}$$

Then, we improve them as follows:

$$\tilde{s} := \mathrm{tr}(\tilde{\mathsf{S}}-\alpha(\mathsf{K}^{-1}\tilde{\mathsf{S}}-\mathsf{I})),$$
$$\bar{s} := \mathrm{tr}(\bar{\mathsf{S}}-\alpha(\mathsf{K}^{-1}\bar{\mathsf{S}}-\mathsf{I})).$$

One can either choose a value for α from the safe range (*e.g.* $\alpha=\frac{2q}{q+2d_{max,avg}}$) or estimate from the samples:

$$\hat{\alpha} = \frac{\widehat{\mathrm{Cov}}(s,\mathrm{tr}(\mathsf{K}^{-1}\bar{\mathsf{S}}-\mathsf{I}))}{\widehat{\mathrm{Var}}(\mathrm{tr}(\mathsf{K}^{-1}\bar{\mathsf{S}}-\mathsf{I}))}.$$

Stratified Sampling

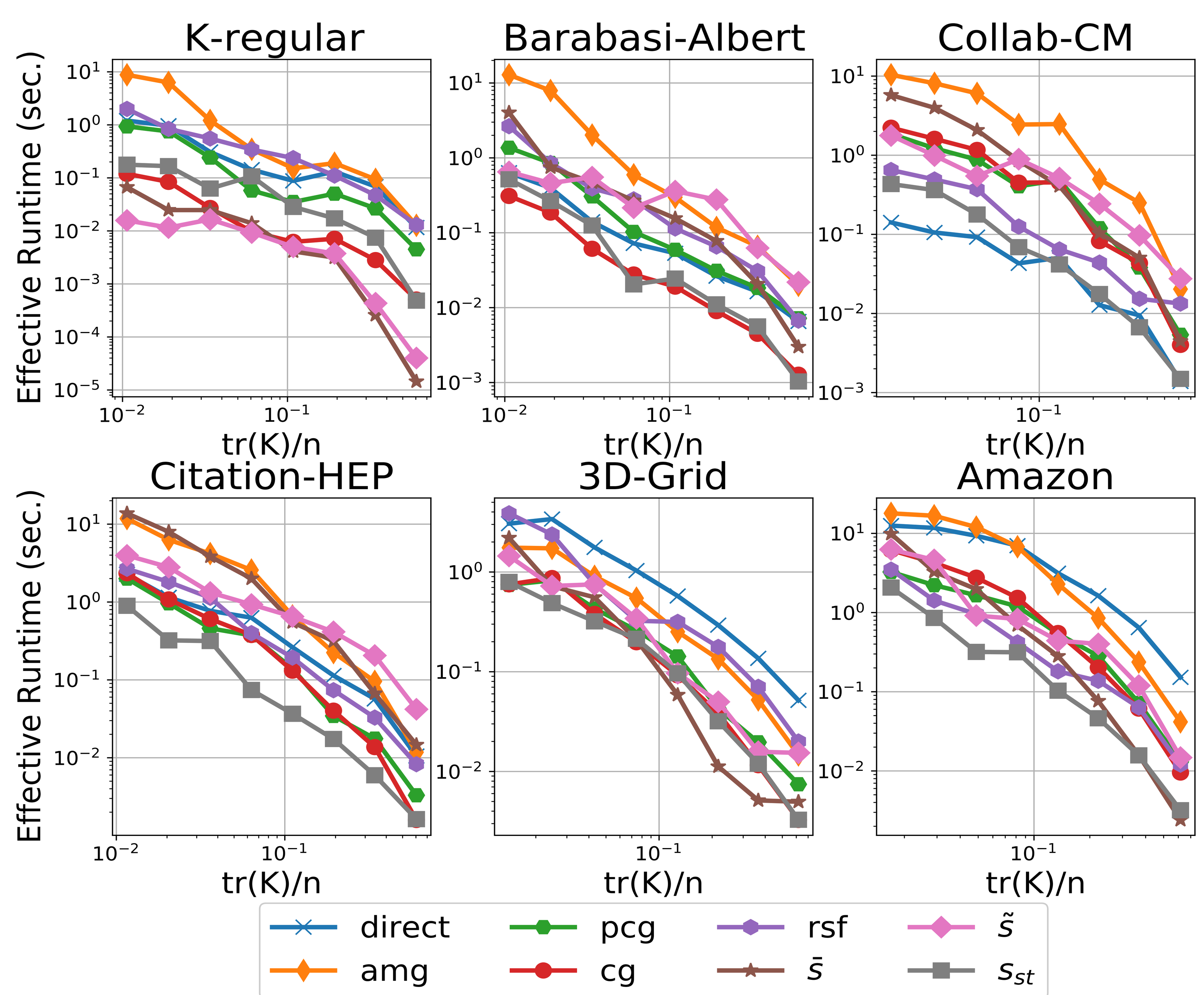
The second method we adapt is stratified sampling.

- Consider a random variable Y with an outcome set $\Omega=\cup_{k=1}^K C_k$.
- We assume:
  - $\mathbb{P}(Y\in C_k)$ is accessible,
  - $s|Y\in C_k$ is easy to sample.
- Then the stratified sampling takes the following form:

$$s_{st} := \sum_{k=1}^{K}\left(\frac{1}{N_k}\sum_{j=1}^{N_k} s^{(j)}|Y\in C_k\right)\mathbb{P}(Y\in C_k).$$

- For certain allocations N_k's, s_{st} has a reduced variance.
- We find such a random variable Y in RSFs!

Comparisons with State-of-the-Art



Grenoble Images Parole Signal Automatique
UMR CNRS 5216 - Grenoble Campus
38400 Saint Martin d'Hères - FRANCE