Minimizing Quadratic Functions in Constant Time

Kohei Hayashi [National Institute of Advanced Industrial Science and Technology] and Yuichi Yoshida [National Institute of Informatics, Preferred Infrastructure, Inc.]

What We Solve

Let $A \in \mathbb{R}^{n\times n}$ be a matrix and $\boldsymbol{d}, \boldsymbol{b} \in \mathbb{R}^n$ be vectors. Then, we consider the following n-dimensional quadratic problem:

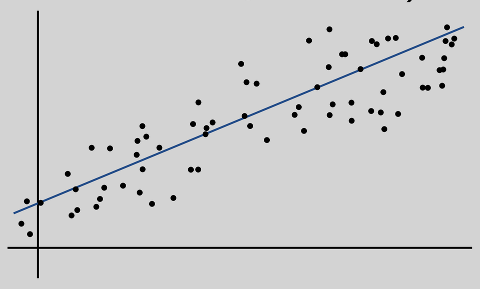
$$z^* = \min_{\boldsymbol{v}\in\mathbb{R}^n} p_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}), \quad (1)$$

$$\text{where } p_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) = \langle \boldsymbol{v}, A\boldsymbol{v}\rangle + n\langle \boldsymbol{v}, \mathrm{diag}(\boldsymbol{d})\boldsymbol{v}\rangle + n\langle \boldsymbol{b}, \boldsymbol{v}\rangle. \quad (2)$$

Caveat: not **argmin** but **min**!

Applications

- Least square distance (a.k.a. linearity check): $\min_{\boldsymbol{w}} \|\boldsymbol{y} - X\boldsymbol{w}\|^2$



- Kernel approximation of the Pearson divergence [Yamada+ NIPS'11]:

$$-\frac{1}{2} - \min_{\boldsymbol{v}\in\mathbb{R}^n} \frac{1}{2}\langle \boldsymbol{v}, H\boldsymbol{v}\rangle - \langle \boldsymbol{h}, \boldsymbol{v}\rangle + \frac{\lambda}{2}\langle \boldsymbol{v}, \boldsymbol{v}\rangle \quad (3)$$

  - $H \in \mathbb{R}^{n\times n}, \boldsymbol{h} \in \mathbb{R}^n$: defined by a kernel function
  - $\lambda \in \mathbb{R}$: regularization coefficient

How Problem (1) Has Been Solved

- Quadratic programming
- Stochastic gradient
- Nyström's method

Problem: All of them need $\Omega(n)$ time

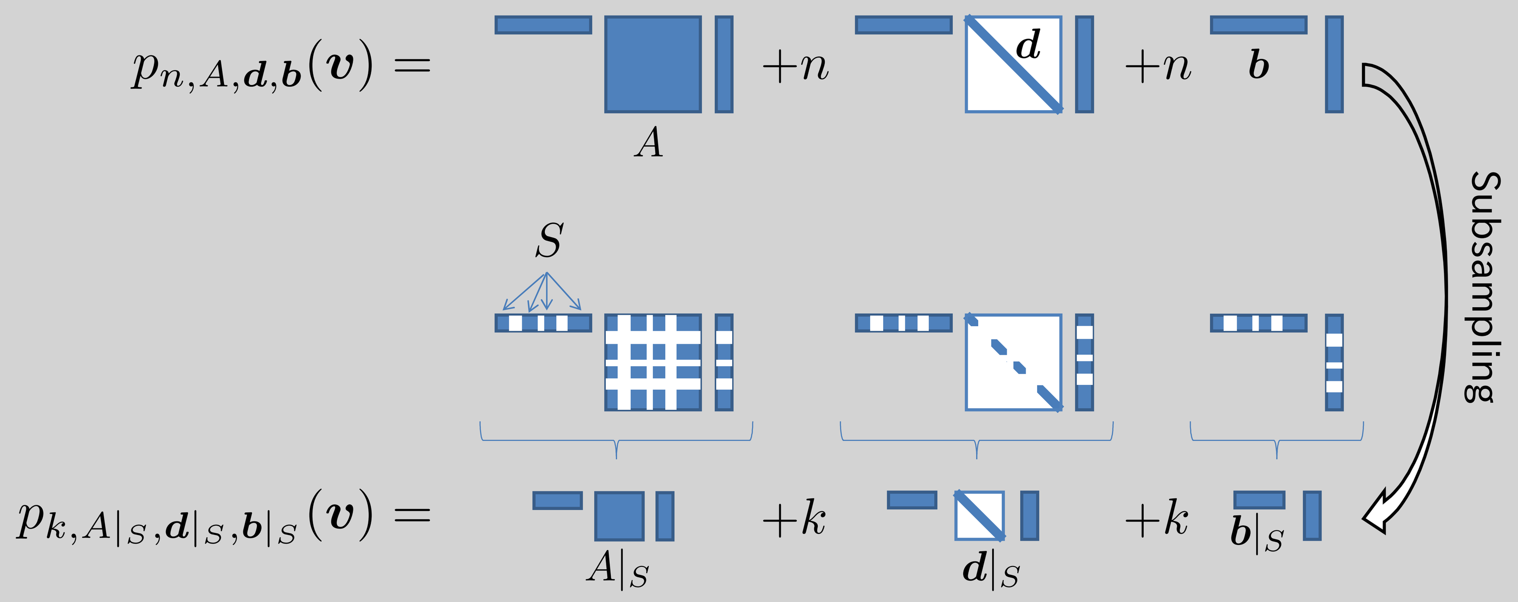
- How can we solve ultra-dimensional problem, e.g. $n \sim 10^{15}$?

Contributions

Goal: Approximately solve (1) in $O(1)$ time

Method: Solve subsampled problem $p_{k,A|_S,\boldsymbol{d}|_S,\boldsymbol{b}|_S}(\boldsymbol{v})$ instead of (1), where

- $k = O(1)$: sampling size
- $S \subset \{1,\ldots,n\}$: sampled indices ($|S| = k$)
- $A|_S \in \mathbb{R}^{k\times k}, \boldsymbol{d}|_S, \boldsymbol{b}|_S \in \mathbb{R}^k$: subsamples of $A, \boldsymbol{d}, \boldsymbol{b}$, resp.



Main Theorem

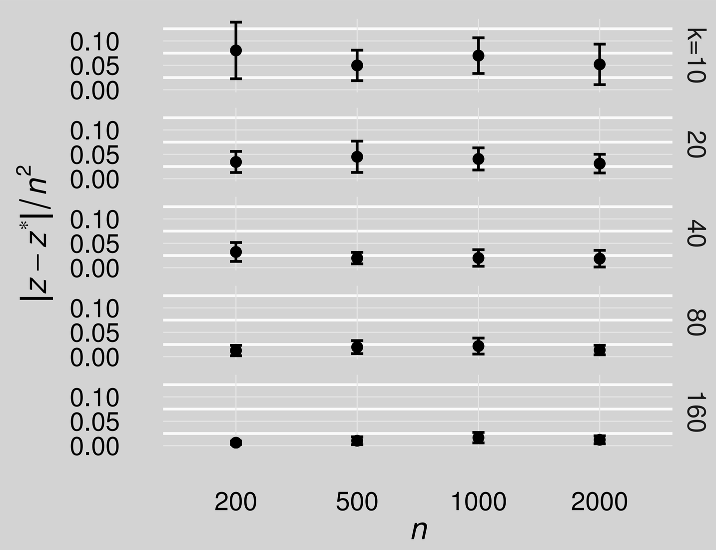
Assume $\forall_{i,j}|A_{ij}|, |b_i|, |d_i| = O(1)$. With parameters $\epsilon, \delta \in (0,1)$, an approximate minimum $z = \frac{n^2}{k^2}\min_{\boldsymbol{v}\in\mathbb{R}^k} p_{k,A|_S,\boldsymbol{d}|_S,\boldsymbol{b}|_S}(\boldsymbol{v})$ in which $k = k(\delta, \epsilon)$ satisfies, with probability at least $1-\delta$,

$$|z - z^*| = O(\epsilon n^2) \quad (4)$$

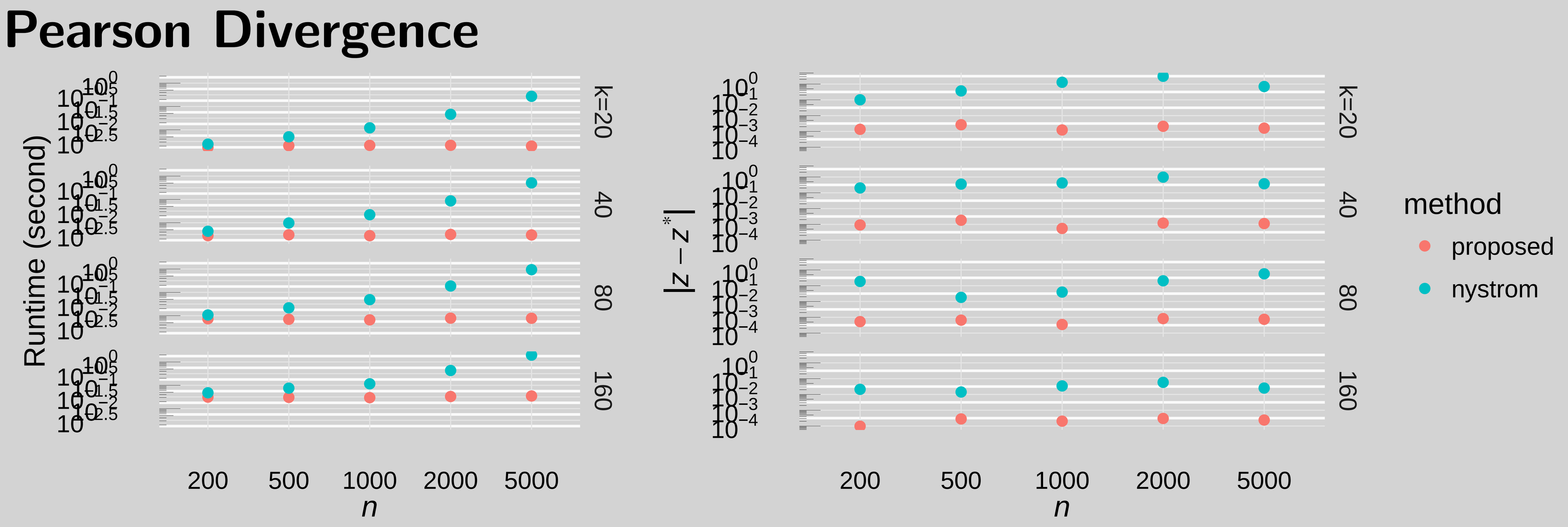
Experiments

Synthetic Data

- $A_{ij}, b_i \sim \mathrm{unif}(-1, 1)$
- $d_i \sim \mathrm{unif}(0, 1)$
- Minimize $p_{k,A|_S,\boldsymbol{d}|_S,\boldsymbol{b}|_S}(\boldsymbol{v})$ by QP



Pearson Divergence



Proof Idea

Rewrite $p_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v})$ as

$$p_{A,D,B}(\boldsymbol{v}) = \langle \boldsymbol{v}, A\boldsymbol{v}\rangle + \langle \boldsymbol{v}^2, D\boldsymbol{1}\rangle + \langle \boldsymbol{v}, B\boldsymbol{1}\rangle,$$

where $(v^2)_i = v_i^2$, $D = \boldsymbol{d}\boldsymbol{1}^\top$, and $B = \boldsymbol{b}\boldsymbol{1}^\top$.

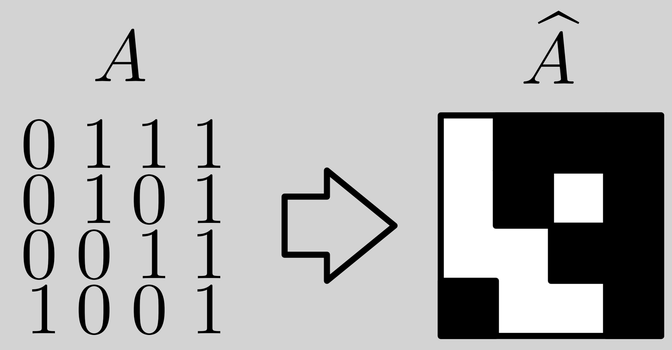
Goal: Show $\min_{\boldsymbol{v}} p_{A,D,B}(\boldsymbol{v}) \approx \frac{n^2}{k^2}\min_{\boldsymbol{v}} p_{A|_S,D|_S,B|_S}(\boldsymbol{v})$.

- Want to say $A \approx A|_S$, $D \approx D|_S$, and $B \approx B|_S$.
- How can we measure the distance between matrices of different sizes?
- Embed matrices to the same space: exploit the graph limit theory.

Dikernel

Dikernel: a measurable function $W : [0,1]^2 \to \mathbb{R}$.

Any matrix $A \in \mathbb{R}^{n\times n}$ has a corresponding dikernel $\widehat{A} : [0,1]^2 \to \mathbb{R}$.



Reduction to Dikernels

For a function $f : [0,1] \to \mathbb{R}$, consider a dikernel analogue of $p_{A,D,B}(f)$:

$$\widehat{p}_{A,D,B}(f) = \langle f, \widehat{A}f\rangle + \langle f^2, \widehat{D}1\rangle + \langle f, \widehat{B}1\rangle,$$

where $\langle f, Wg\rangle = \int_{[0,1]}\int_{[0,1]} W(x,y)f(x)g(y)dxdy$ and $f^2(x) = f(x)^2$.

New goal: Show $\min_f \widehat{p}_{A,D,B}(f) \approx \min_f \widehat{p}_{A|_S,D|_S,B|_S}(f)$.

Key Lemma

Lemma: $|\widehat{p}_{A,D,B}(f) - \widehat{p}_{A|_S,D|_S,B|_S}(f)|$ is small for any bounded $f : [0,1] \to \mathbb{R}$.

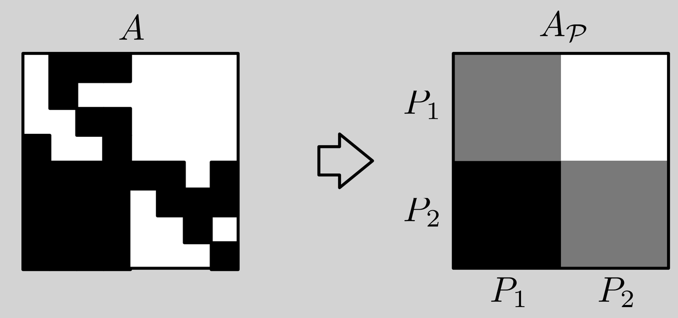
Proof of the new goal: Let $f^* = \mathrm{argmin}_f \widehat{p}_{A,D,B}(f)$ and $f' = \mathrm{argmin}_f \widehat{p}_{A|_S,D|_S,B|_S}(f)$. Then,

$$\widehat{p}_{A,D,B}(f^*) \le \widehat{p}_{A,D,B}(f') \approx \widehat{p}_{A|_S,D|_S,B|_S}(f'),$$
$$\widehat{p}_{A|_S,D|_S,B|_S}(f') \le \widehat{p}_{A|_S,D|_S,B|_S}(f^*) \approx \widehat{p}_{A,D,B}(f^*).$$

Szemerédi's (Weak) Regularity Lemma

Any matrix $A \in \mathbb{R}^{n\times n}$ with $|A_{ij}| = O(1)$ has a partition $\mathcal{P} = (P_1, \ldots, P_k)$ of $\{1, 2, \ldots, n\}$ for constant k with the following property:

Let $A_{\mathcal{P}}$ be the matrix obtained by averaging each part $P_i \times P_j$ of A:



Then, $\|\widehat{A} - \widehat{A_{\mathcal{P}}}\|_\square$ is small.

Cut norm: $\|W\|_\square = \sup_{S,T\subseteq[0,1]} |\int_S\int_T W(x,y)dxdy|$.

Proof of the Key Lemma

Claim: If the cut norm $\|W\|_\square$ is small, then $|\langle f, Wg\rangle|$ is also small.

By the claim,

$$|\widehat{p}_{A,D,B}(f) - \widehat{p}_{A|_S,D|_S,B|_S}(f)|$$
$$\le |\langle f, (\widehat{A} - \widehat{A|_S})f\rangle| + |\langle f^2, (\widehat{D} - \widehat{D|_S})1\rangle| + |\langle f, (\widehat{B} - \widehat{B|_S})1\rangle|$$

is small when $\|\widehat{A} - \widehat{A|_S}\|_\square$, $\|\widehat{D} - \widehat{D|_S}\|_\square$, and $\|\widehat{B} - \widehat{B|_S}\|_\square$ are small.

Let $\mathcal{P}$ be the partition given by Szemerédi's regularity lemma. Then,

$$\|\widehat{A} - \widehat{A|_S}\|_\square \le \|\widehat{A} - \widehat{A_{\mathcal{P}}}\|_\square + \|\widehat{A|_S} - \widehat{A_{\mathcal{P}}}\|_\square,$$

which is small because $A|_S$ has enough information to approximate $A_{\mathcal{P}}$. (The arguments for D and B are almost identical.)