Fastfood - Approximating Kernel Expansions in Loglinear Time

Quoc Le, Tamás Sarlós, Alex Smola. ICML-2013

Zoltán Szabó

Machine Learning Journal Club, Gatsby

May 16, 2014

Notations

- Given: domain ($\mathcal{X}$), kernel k. $\phi : \mathcal{X} \to \mathcal{H}$ feature map

$$k(x, x') = \left\langle \phi(x), \phi(x') \right\rangle_{\mathcal{H}}. \tag{1}$$

- Representer theorem: for many tasks (SVM, ...)

$$w = \sum_{i=1}^{N} \alpha_i \phi(x_i). \tag{2}$$

- Consequence: decision function

$$f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}} = \sum_{i=1}^{N} \alpha_i k(x_i, x). \tag{3}$$

Random kitchen sinks ($\mathcal{X} = \mathbb{R}^d$)

- Bochner Theorem: k continuous, shift invariant $\Leftrightarrow$

$$k(x - x', 0) = \int_{\mathbb{R}^d} e^{-i\langle z, x - x' \rangle} \lambda(z) \mathrm{d}z, \quad \lambda \in \mathcal{M}_+(\mathbb{R}^d) \tag{4}$$

$$= \int_{\mathbb{R}^d} \bar{\phi}_z(x) \phi_z(x') \lambda(z) \mathrm{d}z, \quad \phi_z(x) = e^{izx}. \tag{5}$$

- Assumption: λ is a probability measure (normalization).
- Trick:

$$\hat{k}(x - x', 0) = \frac{1}{n} \sum_{i=1}^{n} e^{-i\left\langle z_j, x - x' \right\rangle}, \quad z_j \sim \lambda. \tag{6}$$

Random kitchen sinks - continued

- Specially, for Gaussians: $k(x-x',0) = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$,

$$\lambda(z) = N\left(z; 0, \frac{I}{\sigma^2}\right), \tag{7}$$

$$k(x-x',0) \approx \left\langle \hat{\phi}(x), \hat{\phi}(x') \right\rangle = \hat{\phi}(x)^* \hat{\phi}(x'), \tag{8}$$

$$\hat{\phi}(x) = \frac{1}{\sqrt{n}} e^{iZx} \in \mathbb{C}^n, \tag{9}$$

$$Z = \left[Z_{ab} \sim N\left(0, \sigma^{-2}\right)\right] \in \mathbb{R}^{n\times d}. \tag{10}$$

- Properties: $\mathcal{O}(nd)$ CPU, $\mathcal{O}(nd)$ RAM.

- Idea (fastfood): do not store Z, only the fast *generators* of $\hat{Z}$.

Fastfood construction: $n = d$ ($d = 2^l$; otherwise padding)

$$V = \frac{1}{\sigma\sqrt{d}} SHGPHB, \tag{11}$$

where

- G: $diag(N(0,1)) \in \mathbb{R}^{d \times d}$.
- P: random permutation matrix $\in \{0,1\}^{d \times d}$.
- B: diag(Bernoulli) $\in \mathbb{R}^{d \times d}$, $B_{ii} \in \{-1,1\}$.
- $H = H_d$: Walsh-Hadamard (WH) transformation $\in \mathbb{R}^{d \times d}$

$$H_1 = 1, H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, H_{2^{k+1}} = \begin{bmatrix} H_{2^k} & H_{2^k} \\ H_{2^k} & -H_{2^k} \end{bmatrix} = (H_2)^{\otimes k}.$$

- S: $\mathrm{diag}(\frac{s_i}{\|G\|_F})$: $s_i \sim \frac{(2\pi)^{\frac{d}{2}} r^{d-1} e^{-\frac{r^2}{2}}}{A_{d-1}}$, $A_{d-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$.

We stack $\frac{n}{d}$ independent copies together:

$$V = [V_1; \ldots; V_{\frac{n}{d}}] = \hat{Z}. \tag{12}$$

Intuition of $V_j = \frac{1}{\sigma\sqrt{d}} SHGPHB$:

- $\frac{1}{\sqrt{d}} HB$: acts as an isometry, which makes the input denser.
- P: ensures the incoherence of the two H-s.
- H, G: WHs with diagonal Gaussian $\approx$ dense Gaussian.
- S: length distributions of V rows are independent.

Fastfood: computational efficiency

- G, B, S:
  - generate them once, store.
  - RAM: $\mathcal{O}(n)$, cost of multiplication: $\mathcal{O}(n)$.
- P: $\mathcal{O}(n)$ storage, $\mathcal{O}(n)$ computation (lookup table).
- H_d: do *not* store,
  - $H_d x$: $\mathcal{O}(d\log(d))$ time/block, $\frac{n}{d}$ blocks $\Rightarrow \mathcal{O}(n\log(d))$.

- To sum up:
  - sinks $\rightarrow$ CPU: $\mathcal{O}(nd)$, RAM: $\mathcal{O}(nd)$, vs
  - fastfood $\rightarrow$ CPU: $\mathcal{O}(n\log(d))$, RAM: $\mathcal{O}(n)$.

Walsh-Hadamard transformation: symmetry, orthogonality

Definition:

$$H_1 = 1, H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, H_{2^{k+1}} = (H_2)^{\otimes k}.$$

Walsh-Hadamard transformation: symmetry, orthogonality

Definition:

$$H_1 = 1, H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, H_{2^{k+1}} = (H_2)^{\otimes k}.$$

Symmetry, orthogonality ($d = 2^k$):

$$H_d = H_d^T, \qquad H_d H_d^T = dI \text{ (i.e., } \frac{1}{\sqrt{d}}H\text{: orth.).} \tag{13}$$

Walsh-Hadamard transformation: symmetry, orthogonality

Definition:

$$H_1 = 1, H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, H_{2^{k+1}} = (H_2)^{\otimes k}.$$

Symmetry, orthogonality ($d = 2^k$):

$$H_d = H_d^T, \qquad H_d H_d^T = dI \text{ (i.e., } \frac{1}{\sqrt{d}}H\text{: orth.).} \tag{13}$$

Proof: H_1, H_2: OK.

$$[H_{2^{k+1}}]^T = \left[(H_2)^{\otimes k}\right]^T = (H_2^T)^{\otimes k} = (H_2)^{\otimes k} = H_{2^{k+1}},$$

$$\begin{aligned} H_{2^k+1} H_{2^{k+1}}^T &= (H_{2^k} \otimes H_2)(H_{2^k} \otimes H_2)^T = (H_{2^k} \otimes H_2)\left(H_{2^k}^T \otimes H_2^T\right) \\ &= \left(H_{2^k} H_{2^k}^T\right) \otimes \left(H_2 H_2^T\right) = (2^k I) \otimes (2I) = 2^{k+1} I \end{aligned}$$

using

$$(A \otimes B)^T = A^T \otimes B^T, \quad (A \otimes B)(C \otimes D) = AC \otimes BD. \tag{14}$$

Walsh-Hadamard transformation: spectral norm

- We have seen ($d = 2^k$): $H_d H_d^T = dI$.
- Spectral norm:

$$\|H_d\|_2 = \sqrt{\lambda_{\max}\left(H_d^T H_d\right)} = \sqrt{\lambda_{\max}(dI)} = \sqrt{d}. \qquad (15)$$

Goal ($\|\cdot\| = \|\cdot\|_2$)

- Unbiasedness:

$$\mathbb{E}\left[\hat{k}(x,x')\right] = \mathbb{E}\left[\hat{\phi}(x)^*\hat{\phi}(x')\right] = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} = k(x,x'). \quad (16)$$

- Concentration:

$$\mathbb{P}\left[\left|\hat{k}(x,x') - k(x,x')\right| \geq a\right] \leq b. \quad (17)$$

Goal – continued

- Low variance (one-block): $v = \frac{x-x'}{\sigma}$, $\psi_j(v) = \cos\left(\frac{[HGPHBv]_j}{\sqrt{d}}\right)$, $j \in [d]$,

$$var[\psi_j(v)] = \frac{1}{2}\left(1 - e^{-\|v\|^2}\right)^2, \tag{18}$$

$$var\left[\sum_{j=1}^{d}\psi_j(v)\right] \leq \frac{d}{2}\left(1 - e^{-\|v\|^2}\right)^2 + dC(\|v\|), \tag{19}$$

$$C(\alpha) = 6\alpha^4\left(e^{-\alpha^2} + \frac{\alpha^2}{3}\right). \tag{20}$$

- Low variance:

$$var\left[\hat{\phi}(x)^T\hat{\phi}(x')\right] \leq \frac{2\left(1 - e^{-\|v\|^2}\right)^2}{n} + \frac{C(\|v\|)}{n}. \tag{21}$$

Proof: $\hat{\phi}(x)^T\hat{\phi}(x') =$ sum of $\frac{n}{d}$ indep. terms $(\times\frac{n}{d})$, averaged $(\times\frac{1}{n^2})$.

Towards unbiasedness: $\mathbb{E}\left([HGPHB]_{ij}\right)$

Let $M := HGPHB$.

$$\mathbb{E}\left(M_{ij}\right) = 0 \tag{22}$$

since

- H_i^T: i^{th} row of $H \Rightarrow H_j$: j^{th} column of H,
- $M_{ij} = (H_i^T)GP(H_j B_{jj})$,
- $M_{ij}|P, B$: sum of independent $N(0,1)$-s, +sign change,
- $\mathbb{E}\left(M_{ij}\right) = \mathbb{E}\left[\mathbb{E}\left(M_{ij}|P,B\right)\right] = \mathbb{E}(0) = 0.$

Unbiasedness: $var\left([HGPHB]_{ij}\right)$

Last slide: $M_{ij} = (H_i^T)GP(H_jB_{jj})$, $\mathbb{E}(M_{ij}) = 0$.

$$\begin{aligned}
var\left(M_{ij}\right) &= \mathbb{E}\left[M_{ij}M_{ij}^T\right] = \mathbb{E}\left[\left(H_i^TGPH_jB_{jj}\right)\left(B_{jj}H_j^TP^TGH_i\right)\right]\\
&= \mathbb{E}\left[B_{jj}^2H_i^TGPee^TP^TGH_i\right] = \mathbb{E}\left[1H_i^TGee^TGH_i\right]\\
&= H_i^T\mathbb{E}\left[G^2\right]H_i = H_i^TIH_i = H_i^TH_i = d
\end{aligned}$$

using $e := [1;\ldots;1] \in \mathbb{R}^d$, $H_jH_j^T = ee^T$, $Pe = e$,
$\mathbb{E}\left(Gee^TG\right) = \mathbb{E}\left(G^2\right)$ (G: diagonal), $E(G_{ii}^2) = 1$.

- We have seen: $\mathbb{E}\left(M_{ij}\right) = 0$, $var\left(M_{ij}\right) = d$.
- $cov(M_{ij}, M_{ik}) = 0$ $(j \neq k)$ since

$$\text{l.h.s.} = \mathbb{E}\left(H_i^T GPH_j B_{jj} H_i^T GPH_k B_{kk}\right) \tag{23}$$

$$= \mathbb{E}\left(B_{jj}B_{kk}\right)\mathbb{E}\left(H_i^T GPH_j H_i^T GPH_k\right), \tag{24}$$

$$\mathbb{E}\left(B_{jj}B_{kk}\right) = \mathbb{E}\left(B_{jj}\right)\mathbb{E}\left(B_{kk}\right) = 0 \times 0 = 0 \tag{25}$$

using that $0 = I\left((B_{jj}, B_{kk}), \text{others}\right) = I(B_{jj}, B_{kk}) = \mathbb{E}(B_{uu})$.

Unbiasedness

In $V = \frac{1}{\sigma\sqrt{d}} HGPHB$ $(V = \frac{1}{\sigma\sqrt{d}} M)$

$$\mathbb{E}\left(V_{ij}\right) = \mathbb{E}\left(\frac{M_{ij}}{\sigma\sqrt{d}}\right) = 0, \tag{26}$$

$$var\left(V_{ij}\right) = var\left(\frac{M_{ij}}{\sigma\sqrt{d}}\right) = \frac{var\left(M_{ij}\right)}{\sigma^2 d} = \frac{d}{\sigma^2 d} = \frac{1}{\sigma^2}, \tag{27}$$

$$cov(V_{ij}, V_{ik}) = 0 \quad (j \neq k). \tag{28}$$

Thus, the distribution of the rows of $V|P, B$: $\sim N\left(0, \frac{I}{\sigma^2}\right)$ $\xrightarrow{\text{[Ali\&Recht 2007]}}$ unbiasedness$|P, B \Rightarrow$ unbiasedness.

Note: we need (i) (28)$|P, B$, but we used $\mathbb{E}_B(B_{jj}B_{kk})$; otherwise: $V \sim$ 'MOG', (ii) the independence of the rows.

Theorem (RBF): Let

$$\hat{k}(x, x') = \frac{1}{d} \sum_{j=1}^{d} \cos\left(\frac{1}{\sigma\sqrt{d}} \left[HGPHB(x - x') \right]_j \right). \qquad (29)$$

Then

$$\mathbb{P}\left[\left| \hat{k}(x, x') - k(x, x') \right| \geq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{d}} \alpha \right] \leq 2\delta \qquad (30)$$

for $\delta > 0$, $\alpha = \frac{2\|x - x'\|}{\sigma} \sqrt{\log\left(\frac{2d}{\delta}\right)}$.

Concentration – proof

- We have already seen: $\mathbb{E}\left[\hat{k}(x,x')\right] = k(x,x')$.
- Lemma (concentration of Gaussian measure; Ledoux 1996): $f:\mathbb{R}^d \to \mathbb{R}$ Lipschitz continuous (L), $g \sim N(0, I_d)$. Then

$$\mathbb{P}\left(|f(g) - \mathbb{E}[f(g)]| \geq t\right) \leq 2e^{-\frac{t^2}{2L^2}}. \tag{31}$$

- Lemma [approximate isometry of $\frac{HB}{\sqrt{d}}$; Ailon & Chazelle, 2009]: $x \in \mathbb{R}^d$; H, B: from V. For any $\delta > 0$

$$\mathbb{P}\left[\left\|\frac{HBx}{\sqrt{d}}\right\|_\infty \geq \|x\|_2 \sqrt{\log\left(\frac{2d}{\delta}\right)\frac{2}{d}}\right] \leq \delta. \tag{32}$$

Concentration – proof

- Notation: $v = \frac{x-x'}{\sigma}$, $k(v) = k(x,x')$, $\hat{k}(v) = \hat{k}(x,x')$.
- Sufficient to prove:

$$f(G,P,B) = \frac{1}{d}\sum_{j=1}^{d}\cos(z_j), \quad z = HGu, \quad u = P\frac{HB}{\sqrt{d}}v \quad (33)$$

  concentrates around the mean.
- Idea:
  - $G \mapsto f(G,P,B)$: Lipschitz $\Rightarrow$ high-$\mathbb{P}$ concentration in $G|B$.
  - Approximate isometry of $\frac{HB}{\sqrt{d}}$: high-$\mathbb{P}$ in B (P: does not matter).
  - Union bound.

Concentration – proof

$$h(a) = \frac{1}{d}\sum_{j=1}^{d}\cos(a_j) \quad (a \in \mathbb{R}^d),$$

$$\left|f(G;P,B) - f(G';P,B)\right| = \left|h[Hdiag(g)u] - h[Hdiag(g')u]\right|,$$

$$|h(a) - h(b)| = \frac{1}{d}\left|\sum_{j=1}^{d}\cos(a_j) - \cos(b_j)\right|$$

$$\leq \frac{1}{d}\sum_{j=1}^{d}|\cos(a_j) - \cos(b_j)| \leq \frac{1}{d}\sum_{j=1}^{d}|a_j - b_j|$$

$$= \frac{1}{d}\|a-b\|_1 \leq \frac{1}{d}\sqrt{d}\|a-b\|_2 = \frac{1}{\sqrt{d}}\|a-b\|_2$$

$$\left\|Hdiag(g)u - Hdiag(g')u\right\|_2 \leq \|H\|_2\left\|diag(g-g')u\right\|_2$$

$$= \sqrt{d}\left\|(g-g')\circ u\right\|_2 \leq \sqrt{d}\left\|g-g'\right\|_2\|u\|_\infty.$$

Concentration – proof

Until now:

$$\left|f(G;P,B) - f(G';P,B)\right| \leq \|u\|_\infty \left\|g - g'\right\|_2. \tag{34}$$

$\|u\|_\infty$ term: using $\|Pw\|_\infty = \|w\|_\infty$

$$\left\|u = P\frac{HB}{\sqrt{d}}v\right\|_\infty = \left\|\frac{HB}{\sqrt{d}}v\right\|_\infty. \tag{35}$$

Approximate isometry of $\frac{HB}{\sqrt{d}}$: with $1-\delta$ $\mathbb{P}_{B,P}$-probability

$$\|u\|_\infty \leq \|v\|_2 \sqrt{\log\left(\frac{2d}{\delta}\right)\frac{2}{d}}. \tag{36}$$

Until now: f Lipschitz with $1-\delta$ $\mathbb{P}_{B,P}$-probability

$$\left|f(G;P,B)-f(G';P,B)\right| \leq \left[\|v\|_2\sqrt{\log\left(\frac{2d}{\delta}\right)\frac{2}{d}}\right]\left\|g-g'\right\|_2$$
$$=: L\left\|g-g'\right\|_2.$$

By the concentration of the Gaussian measure $[G_{ii}\sim N(0,1)]$:

$$\mathbb{P}_G\left[|f(G;P,B)-k(v)|\geq t\right]\leq 2e^{\frac{-t^2}{2L^2}} =: \delta, \quad (37)$$

$$\mathbb{P}_G\left[|f(G;P,B)-k(v)|\geq\sqrt{2\log\left(\frac{2}{\delta}\right)}L\right]\leq\delta. \quad (38)$$

We apply a union bound: $\Rightarrow 2\delta$.

Low variance: $var[\psi_j(v)]$

- Notations:

$$w = \frac{1}{\sqrt{d}} HBv, u = Pw, z = HGu. \tag{39}$$

- High-level idea:
  - $cov(z_j, z_t|u)$: normal.
  - $cov(\psi(z_j), \psi(z_t)|u)$, some $exp - cosh$ relations, $j = t$.

Low variance: $z_j|u$

Def.: $w = \frac{1}{\sqrt{d}}HBv$, $u = Pw$, $z = HGu$. Using $\mathbb{E}_G(HGu|u) = 0$

$$\begin{aligned}
cov(z_j, z_j|u) &= cov([HGu]_j, [HGu]_j|u) = cov(H_j^T Gu, H_j^T Gu|u) \\
&= \mathbb{E}\left[\left(H_j^T Gu\right)\left(H_j^T Gu\right)^T\right], \\
H_j^T Gu &= [H_{j1}G_{11}u_1, H_{j2}G_{22}u_2, \ldots], \quad (G : \text{diagonal}) \\
cov(z_j, z_j|u) &= \mathbb{E}\left(\sum_i G_{ii}^2 H_{ji}^2 u_i^2\right) = \sum_i \mathbb{E}\left(G_{ii}^2\right) u_i^2 = \sum_i u_i^2 \\
&= \|u\|^2 = \|v\|^2
\end{aligned}$$

using $H_{ji}^2 = 1$ $(H_{ji} = \pm 1)$, $\mathbb{E}\left(G_{ii}^2\right) = 1$ $[G_{ii} \sim N(0,1)]$, isometry of $P\frac{1}{\sqrt{d}}HB$.$\}$ $\Rightarrow$ $z|u$: normal, $z_j|u \sim N\left(0, \|v\|^2\right)$.

Low variance: $cov(z_j, z_t|u)$

Last slide: $z_j|u \sim N\left(0, \|v\|^2\right)$.

$$cov(z_j, z_t|u) = corr(z_j, z_t|u) std(z_j|u) std(z_t|u) \tag{40}$$

$$= corr(z_j, z_t|u) \|v\|^2 =: \rho_{jt}(u) \|v\|^2 =: \rho \|v\|^2 . \tag{41}$$

$$\begin{bmatrix} z_j \\ z_t \end{bmatrix} \sim N\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \|v\|^2 =: LL^T\right) = N(0, Lg), g \sim N(0, I) \tag{42}$$

$$L = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \|v\| . \tag{43}$$

Now for $\psi_j(v) = \cos(z_j)$:

$$cov(\psi_j(v), \psi_t(v)|u) = cov(\cos([Lg]_1), \cos([Lg]_2)) \tag{44}$$

$$= \mathbb{E}_g \left[\prod_{k=1}^{2} \cos([Lg]_k)\right] - \prod_{k=1}^{2} \mathbb{E}_g \left[\cos([Lg]_k)\right] . \tag{45}$$

Low variance: first term in $cov(\psi_j(v), \psi_t(v)|u)$

Using $\cos(\alpha)\cos(\beta) = \frac{1}{2}[\cos(\alpha-\beta)+\cos(\alpha+\beta)]$, $g = [g_1; g_2]$,

$$\mathbb{E}_g\left[\cos([Lg]_1)\cos([Lg]_2)\right] = \frac{1}{2}\mathbb{E}_g\left\{\cos([Lg]_1 - [Lg]_2) + \cos([Lg]_1 + [Lg]_2)\right\},$$

where

$$[Lg]_1 - [Lg]_2 = \|v\|\left(g_1 - \rho g_1 - \sqrt{1-\rho^2}g_2\right) = \|v\|\sqrt{2-2\rho}h,$$
$$[Lg]_1 + [Lg]_2 = \|v\|\left(g_1 + \rho g_1 + \sqrt{1-\rho^2}g_2\right) = \|v\|\sqrt{2+2\rho}h$$

since

$$g_1 - \rho g_1 - \sqrt{1-\rho^2}g_2 \sim \sqrt{(1-\rho)^2 + (1-\rho^2)}h = \sqrt{2-2\rho}h,$$
$$g_1 + \rho g_1 + \sqrt{1-\rho^2}g_2 \sim \sqrt{(1+\rho)^2 + (1-\rho^2)}h = \sqrt{2+2\rho}h.$$

where $h \sim N(0,1)$.

Thus

$$\mathbb{E}_g\left[\cos([Lg]_1)\cos([Lg]_2)\right] = \frac{1}{2}\mathbb{E}_g\left\{\cos(a_- h) + \cos(a_+ h)\right\}, \tag{46}$$

$$a_- = \|v\|\sqrt{2-2\rho}, \tag{47}$$

$$a_+ = \|v\|\sqrt{2+2\rho}. \tag{48}$$

Making use of the relation

$$\mathbb{E}[\cos(ah)] = e^{-\frac{1}{2}a^2}, \quad h \sim N(0,1), \tag{49}$$

we obtained

$$\mathbb{E}_g\left[\cos([Lg]_1)\cos([Lg]_2)\right] = \frac{1}{2}\left[e^{-\|v\|^2(1-\rho)} + e^{-\|v\|^2(1+\rho)}\right].$$

Low variance: value of $\mathbb{E}[\cos(b)]$

Lemma:

$$\mathbb{E}[\cos(b)] = e^{-\frac{1}{2}\sigma^2}, \quad h \sim N(0,\sigma^2), \tag{50}$$

Proof: The characteristic function of $b \sim N(m,\sigma^2)$

$$c(t) = \mathbb{E}_b\left[e^{jtb}\right] = e^{itm-\frac{1}{2}\sigma^2 t^2}. \tag{51}$$

Specially, for $m = 0$, $t = 1$ ($b \sim N(0,\sigma^2)$)

$$e^{-\frac{1}{2}\sigma^2} = \mathbb{E}_b\left[e^{jb}\right] = \mathbb{E}\left[cos(b)\right]. \tag{52}$$

Since $z_j \sim N(0, \|v\|^2)$

$$\mathbb{E}_g[\cos(z_j)]\mathbb{E}_g[\cos(z_t)] = (\mathbb{E}_g\left[\cos(\|v\|\, h)\right])^2 = \left(e^{-\frac{1}{2}\|v\|^2}\right)^2 = e^{-\|v\|^2}$$

using the identity for $\mathbb{E}[\cos(ah)]$. Thus $[cosh(a) = \frac{e^a+e^{-a}}{2}]$

$$\begin{aligned} cov(\psi_j(v), \psi_t(v)|u) &= \frac{1}{2}\left[e^{-\|v\|^2(1-\rho)} + e^{-\|v\|^2(1+\rho)}\right] - e^{-\|v\|^2} \\ &= e^{-\|v\|^2}\left[\frac{e^{\|v\|^2\rho} + e^{-\|v\|^2\rho}}{2} - 1\right] \\ &= e^{-\|v\|^2}\left[cosh\left(\|v\|^2\rho\right) - 1\right]. \end{aligned}$$

Low variance: $var[\psi_j(v)]$

With $j = t$, $\rho = 1$ we got

$$var[\psi_j(v)] = e^{-\|v\|^2} \left[\frac{e^{\|v\|^2} + e^{-\|v\|^2}}{2} - 1 \right] \tag{53}$$

$$= \frac{1 + e^{-2\|v\|^2}}{2} - e^{-\|v\|^2} \tag{54}$$

$$= \frac{1}{2} \left(1 - 2e^{-\|v\|^2} + e^{-2\|v\|^2} \right) \tag{55}$$

$$= \frac{1}{2} \left(1 - e^{-\|v\|^2} \right)^2 . \tag{56}$$

Low variance: $var\left[\sum_{j=1}^{d}\psi_j(v)\right]$

Decomposition:

$$var\left[\sum_{j=1}^{d}\psi_j(v)\right] = \sum_{j,t=1}^{d} cov\left[\psi_j(v), \psi_t(v)\right]. \tag{57}$$

We have seen that

$$cov\left[\psi_j(v), \psi_t(v)|u\right] = e^{-\|v\|^2}\left[cosh\left(\|v\|^2\rho\right) - 1\right]. \tag{58}$$

We rewrite the *cosh* term.

Low variance: $cosh\left(\|v\|^2\rho\right)$

Third-order Taylor expansion around 0 with remainder term

$$cosh\left(\|v\|^2\rho\right) = 1 + \frac{1}{2!}\|v\|^4\rho^2 + \frac{1}{3!}sinh(\eta)\|v\|^6\rho^3 \tag{59}$$
$$\leq 1 + \frac{1}{2}\|v\|^4\rho^2 + \frac{1}{6}sinh\left(\|v\|^2\right)\|v\|^6\rho^3 \tag{60}$$
$$\leq 1 + \|v\|^4\rho^2 B(\|v\|), \tag{61}$$

where

- $\eta \in \left[-\|v\|^2|\rho|, \|v\|^2|\rho|\right]$,
- we used: $cosh' = sinh$, $sinh' = cosh$, $cosh(0) = \frac{1}{2}$, $sinh(a) = \frac{e^a - e^{-a}}{2}$, $sinh(0) = 0$, monotonicity of $sinh$, $|\rho| \leq 1$.
- $B(\|v\|) = \frac{1}{2} + \frac{1}{6}sinh\left(\|v\|^2\right)\|v\|^2$, $(\rho^3 \leq \rho^2)$.

- Plugging the result back to $cov\left[\psi_j(v),\psi_t(v)|u\right]$, $e^{-\|v\|^2}\le 1$:
$$cov\left[\psi_j(v),\psi_t(v)|u\right]\le\|v\|^4\rho^2 B(\|v\|). \tag{62}$$
Here, $\rho=\rho(u)$.
- Remains: to bound $\mathbb{E}_u\left[\rho^2(u)\right]$.
- Small if $\mathbb{E}\left(\|u\|_4^4\right)$ is small ($\Leftarrow$ *HB*: randomized preconditioner).

Numerical experiments

- Accuracy: similar to random kitchen sinks (RKS).
- CPU, RAM:

d	n	Fastfood	RKS	Speedup	RAM
$1,024$	$16,384$	0.00058s	0.0139s	24x	256x
$4,096$	$32,768$	0.00136s	0.1224s	90x	1024x
$8,192$	$65,536$	0.00268s	0.5360s	200x	2048x

Summary

- Random kitchen sinks: use
  - (normally distributed) random projections, which
  - are stored (Z).
- Fastfood:
  - approximates the RKS features using the composition of
  - diag, permutation, Walsh-Hadamard transformations ($\hat{Z}$).
  - *does not store the feature map*!
- Results:
  - unbiased, concentration, low variance,
  - RAM + CPU improvements.

Let $M = HGPHB$. Squared norm of the j^{th} row

$$l_j^2 = \left[MM^T\right]_{jj} = \left[(HGPHB)(HGPHB)^T\right]_{jj} \tag{63}$$
$$= \left[HGPHBB^T H^T P^T G H^T\right]_{jj} = \left[dHG^2H^T\right]_{jj} \tag{64}$$
$$= d\sum_i H_{ij}^2 G_{ii}^2 = d\sum_i G_{ii}^2 = d\,\|G\|_F^2 \tag{65}$$

by $BB^T = I$ [B=diag(± 1)], $HH^T = dI$, $PP^T = I$, $H_{ij}^2 = 1$ ($H_{ij} = \pm 1$).

Fastfood: optional scaling matrix (S)

- Previous slide: $l_j^2 = d \left\| G \right\|_F^2$.
- Rescaling by $\frac{1}{l_j} = \frac{1}{\sqrt{d} \left\| G \right\|_F}$: yields rows of unit length.
- S:
  - $\operatorname{diag}\left(\frac{s_i}{\left\| G \right\|_F}\right)$: $s_i \sim \frac{(2\pi)^{\frac{d}{2}} r^{d-1} e^{-\frac{r^2}{2}}}{A_{d-1}}$, $A_{d-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$.
  - length distributions of the V rows: independent of each other.