

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

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

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

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

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

- Consequence: decision function

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

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

- Bochner Theorem: k continuous, shift invariant \Leftrightarrow

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

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

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

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

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

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

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

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

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

- 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)

$$V = \frac{1}{\sigma\sqrt{d}} SHGP HB, \quad (11)$$

where

- G : $\text{diag}(N(0, 1)) \in \mathbb{R}^{d \times d}$.
- P : random permutation matrix $\in \{0, 1\}^{d \times d}$.
- B : $\text{diag}(\text{Bernoulli}) \in \mathbb{R}^{d \times d}$, $B_{ij} \in \{-1, 1\}$.
- $H = H_d$: 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 : $\text{diag}(\frac{s_j}{\|G\|_F})$: $s_j \sim \frac{(2\pi)^{\frac{d}{2}} r^{d-1} e^{-\frac{r^2}{2}}}{A_{d-1}}$, $A_{d-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$.

Fastfood construction: $n > d$ (assumption: $d|n$)

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

$$V = [V_1; \dots; V_{\frac{n}{d}}] = \hat{Z}. \quad (12)$$

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 ($d = 2^k$):

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

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 ($d = 2^k$):

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

Proof: H_1, H_2 : OK.

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

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

using

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

Walsh-Hadamard transformation: spectral norm

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

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

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

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

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

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

- Low variance:

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

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

Let $M := HGPHB$.

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

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,
- $\mathbb{E}(M_{ij}) = \mathbb{E}[\mathbb{E}(M_{ij}|P, B)] = \mathbb{E}(0) = 0$.

Unbiasedness: $\text{var}([HGP]_{ij})$

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

$$\begin{aligned}\text{var}(M_{ij}) &= \mathbb{E} \left[M_{ij} M_{ij}^T \right] = \mathbb{E} \left[\left(H_i^T G P H_j B_{jj} \right) \left(B_{jj} H_j^T P^T G H_i \right) \right] \\ &= \mathbb{E} \left[B_{jj}^2 H_i^T G P e e^T P^T G H_i \right] = \mathbb{E} \left[1 H_i^T G e e^T G H_i \right] \\ &= H_i^T \mathbb{E} \left[G^2 \right] H_i = H_i^T I H_i = H_i^T H_i = d\end{aligned}$$

using $e := [1; \dots; 1] \in \mathbb{R}^d$, $H_j H_j^T = e e^T$, $P e = e$,
 $\mathbb{E}(G e e^T G) = \mathbb{E}(G^2)$ (G : diagonal), $E(G_{ii}^2) = 1$.

Unbiasedness: $\text{cov}([HGPHB]_{ij}, [HGPHB]_{ik}), j \neq k$

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

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

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

$$\mathbb{E}(B_{jj} B_{kk}) = \mathbb{E}(B_{jj}) \mathbb{E}(B_{kk}) = 0 \times 0 = 0 \quad (25)$$

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

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

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

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

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

Thus, the distribution of the rows of $V|P, B: \sim N(0, \frac{1}{\sigma^2})$

[Ali&Recht 2007] \rightarrow 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.

Concentration ($e \rightarrow \cos$, $n = d$)

Theorem (RBF): Let

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

Then

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

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

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

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

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

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

- 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.

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

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

$$\begin{aligned} |h(a) - h(b)| &= \frac{1}{d} \left| \sum_{j=1}^d \cos(a_j) - \cos(b_j) \right| \\ &\leq \frac{1}{d} \sum_{j=1}^d |\cos(a_j) - \cos(b_j)| \leq \frac{1}{d} \sum_{j=1}^d |a_j - b_j| \\ &= \frac{1}{d} \|a - b\|_1 \leq \frac{1}{d} \sqrt{d} \|a - b\|_2 = \frac{1}{\sqrt{d}} \|a - b\|_2 \end{aligned}$$

$$\begin{aligned} \|Hdiag(g)u - Hdiag(g')u\|_2 &\leq \|H\|_2 \|diag(g - g')u\|_2 \\ &= \sqrt{d} \|(g - g') \circ u\|_2 \leq \sqrt{d} \|g - g'\|_2 \|u\|_\infty. \end{aligned}$$

Until now:

$$|f(G; P, B) - f(G'; P, B)| \leq \|u\|_\infty \|g - g'\|_2. \quad (34)$$

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

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

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

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

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

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

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

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

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

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

- Notations:

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

- High-level idea:

- $\text{cov}(z_j, z_t|u)$: normal.
- $\text{cov}(\psi(z_j), \psi(z_t)|u)$, some *exp – cosh* relations, $j = t$.

Low variance: $z_j|u$

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

$$\begin{aligned} \text{cov}(z_j, z_j|u) &= \text{cov}([HGu]_j, [HGu]_j|u) = \text{cov}(H_j^T Gu, H_j^T Gu|u) \\ &= \mathbb{E} \left[\left(H_j^T Gu \right) \left(H_j^T Gu \right)^T \right], \end{aligned}$$

$$H_j^T Gu = [H_{j1} G_{11} u_1, H_{j2} G_{22} u_2, \dots], \quad (G : \text{diagonal})$$

$$\begin{aligned} \text{cov}(z_j, z_j|u) &= \mathbb{E} \left(\sum_i G_{ii}^2 H_{ji}^2 u_i^2 \right) = \sum_i \mathbb{E} (G_{ii}^2) u_i^2 = \sum_i u_i^2 \\ &= \|u\|^2 = \|v\|^2 \end{aligned}$$

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

Low variance: $\text{cov}(z_j, z_t | u)$

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

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

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

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

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

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

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

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

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

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

$$\mathbb{E}_{\mathbf{g}} [\cos([L\mathbf{g}]_1) \cos([L\mathbf{g}]_2)] = \frac{1}{2} \mathbb{E}_{\mathbf{g}} \{ \cos([L\mathbf{g}]_1 - [L\mathbf{g}]_2) + \cos([L\mathbf{g}]_1 + [L\mathbf{g}]_2) \},$$

where

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

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

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

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

$$a_- = \|\mathbf{v}\| \sqrt{2 - 2\rho}, \quad (47)$$

$$a_+ = \|\mathbf{v}\| \sqrt{2 + 2\rho}. \quad (48)$$

Making use of the relation

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

we obtained

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

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

Lemma:

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

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

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

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

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

Low variance: second term in $\text{cov}(\psi_j(v), \psi_t(v)|u)$

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

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

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

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

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

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

$$= \frac{1 + e^{-2\|\mathbf{v}\|^2}}{2} - e^{-\|\mathbf{v}\|^2} \quad (54)$$

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

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

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

Decomposition:

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

We have seen that

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

We rewrite the *cosh* term.

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

Third-order Taylor expansion around 0 with remainder term

$$\cosh\left(\|v\|^2 \rho\right) = 1 + \frac{1}{2!} \|v\|^4 \rho^2 + \frac{1}{3!} \sinh(\eta) \|v\|^6 \rho^3 \quad (59)$$

$$\leq 1 + \frac{1}{2} \|v\|^4 \rho^2 + \frac{1}{6} \sinh\left(\|v\|^2\right) \|v\|^6 \rho^3 \quad (60)$$

$$\leq 1 + \|v\|^4 \rho^2 B(\|v\|), \quad (61)$$

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 $\text{cov} [\psi_j(v), \psi_t(v)|u]$, $e^{-\|v\|^2} \leq 1$:

$$\text{cov} [\psi_j(v), \psi_t(v)|u] \leq \|v\|^4 \rho^2 B(\|v\|). \quad (62)$$

Here, $\rho = \rho(u)$.

- Remains: to bound $\mathbb{E}_u [\rho^2(u)]$.
- 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

- 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.

Fastfood: properties - rows of $HGPHB$: same length

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

$$l_j^2 = [MM^T]_{jj} = [(HGPHB)(HGPHB)^T]_{jj} \quad (63)$$

$$= [HGPHB B^T H^T P^T G H^T]_{jj} = [dHG^2 H^T]_{jj} \quad (64)$$

$$= d \sum_i H_{ij}^2 G_{ii}^2 = d \sum_i G_{ii}^2 = d \|G\|_F^2 \quad (65)$$

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

Fastfood: optional scaling matrix (S)

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