A Proofs

A.1 Proof of Theorem 4.1

Proof. By the definition of $\langle \nu_{\mathrm{DE}}, f^{(1)}_{\mathrm{DE}}, f^{(2)}_{\mathrm{DE}} \rangle$, we have:

$$\begin{aligned}
\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y}) &= (2\pi)^{-\frac{d}{2}} D^2 \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2}\|\boldsymbol{\omega}\|^2 + 2\boldsymbol{\omega}^\top \mathbf{A}\boldsymbol{\omega} \\
&+ \boldsymbol{\omega}^\top (\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}) + \mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y} \Big) d\boldsymbol{\omega} \\
&= (2\pi)^{-\frac{d}{2}} D^2 \exp\left(\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y} \right) \\
&\times \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2}\boldsymbol{\omega}^\top (\mathbf{I}_d - 4\mathbf{A})\boldsymbol{\omega} + \boldsymbol{\omega}^\top (\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}) \Big) d\boldsymbol{\omega}.
\end{aligned}$$

Since $8\mathbf{A} \prec \mathbf{I}_d$, we have $4\mathbf{A} \prec 0.5\mathbf{I}_d \prec \mathbf{I}_d$, meaning that $\mathbf{I}_d - 4\mathbf{A}$ is positive definite and invertible. The following identity is straightforward to check:

$$\begin{gathered}
-\frac{1}{2}\boldsymbol{\omega}^\top (\mathbf{I}_d - 4\mathbf{A})\boldsymbol{\omega} + \boldsymbol{\omega}^\top (\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}) = -\frac{1}{2}\left(\boldsymbol{\omega} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \\
\boldsymbol{\Sigma} = (\mathbf{I}_d - 4\mathbf{A})^{-1}, \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}).
\end{gathered}$$

Therefore, we have:

$$\begin{aligned}
\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y}) &= (2\pi)^{-d/2} D^2 \exp\left(\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y} + \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \right) \\
&\times \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left(\boldsymbol{\omega} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}) \right) d\boldsymbol{\omega}.
\end{aligned}$$

Next, we use the fact that the integral of the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ is 1:

$$(2\pi)^{-d/2} \det(\boldsymbol{\Sigma})^{-1/2} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left(\boldsymbol{\omega} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}) \right) d\boldsymbol{\omega} = 1.$$

From that we conclude:

$$\begin{aligned}
\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y}) &= D^2 \det(\boldsymbol{\Sigma})^{1/2} \exp\left(\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y} + \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \right) \\
&= D^2 \det(\mathbf{I}_d - 4\mathbf{A})^{-1/2} \exp\Big(\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + \mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y} \\
&+ \frac{1}{2}(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y})^\top (\mathbf{I}_d - 4\mathbf{A})^{-1}(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y}) \Big) \\
&= D^2 \det(\mathbf{I}_d - 4\mathbf{A})^{-1/2} \exp\Big(\mathbf{x}^\top \left(\mathbf{C}^{(1)} + \frac{1}{2}(\mathbf{B}^{(1)})^\top (\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B}^{(1)} \right) \mathbf{x} \\
&+ \mathbf{y}^\top \left(\mathbf{C}^{(2)} + \frac{1}{2}(\mathbf{B}^{(2)})^\top (\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B}^{(2)} \right) \mathbf{y} + \mathbf{x}^\top (\mathbf{B}^{(1)})^\top (\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B}^{(2)}\mathbf{y} \Big).
\end{aligned}$$

Based on this expression, we conclude that, indeed, $\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y}) = K^{(0)}(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ if the conditions from theorem's statement are satisfied.

Next, we calculate expression for the variance. For any random variable Z, $\operatorname{Var} Z = \mathbb{E}Z^2 - (\mathbb{E}Z)^2$. In particular, if $Z = f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y})$, $\boldsymbol{\omega} \sim \nu_{\mathrm{DE}}$, we get:

$$\begin{aligned}
\operatorname{Var}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y}) &= \mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x})^2 f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y})^2 - \left(\mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y}) \right)^2 \\
&= \mathbb{E}_{\nu_{\mathrm{DE}}} f^{(1)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{x})^2 f^{(2)}_{\mathrm{DE}}(\boldsymbol{\omega}, \mathbf{y})^2 - K^{(0)}(\mathbf{x}, \mathbf{y})^2.
\end{aligned}$$

We have:

$$\mathbb{E}_{\nu_{\mathrm{DE}}} f_{\mathrm{DE}}^{(1)}(\boldsymbol{\omega}, \mathbf{x})^2 f_{\mathrm{DE}}^{(2)}(\boldsymbol{\omega}, \mathbf{y})^2 = (2\pi)^{\frac{d}{2}} D^4 \int_{\mathbb{R}^d} \exp\Bigg(-\frac{1}{2}\|\boldsymbol{\omega}\|^2 + 4\boldsymbol{\omega}^\top \mathbf{A}\boldsymbol{\omega} + 2\boldsymbol{\omega}^\top(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y})$$
$$+ 2\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + 2\mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y}\Bigg) d\boldsymbol{\omega} = (2\pi)^{\frac{d}{2}} D^4 \exp\Big(2\mathbf{x}^\top \mathbf{C}^{(1)}\mathbf{x} + 2\mathbf{y}^\top \mathbf{C}^{(2)}\mathbf{y}\Big)$$
$$\times \int_{\mathbb{R}^d} \exp\Bigg(-\frac{1}{2}\boldsymbol{\omega}^\top(\mathbf{I}_d - 8\mathbf{A})\boldsymbol{\omega} + 2\boldsymbol{\omega}^\top(\mathbf{B}^{(1)}\mathbf{x} + \mathbf{B}^{(2)}\mathbf{y})\Bigg) d\boldsymbol{\omega}.$$

Evaluation of the integral above can be done in the same way as calculation of $\mathbb{E}_{\nu_{\mathrm{DE}}} f_{\mathrm{DE}}^{(1)}(\boldsymbol{\omega}, \mathbf{x}) f_{\mathrm{DE}}^{(2)}(\boldsymbol{\omega}, \mathbf{y})$, noticing that $\mathbf{I}_d - 8\mathbf{A}$ is positive definite and invertible. The result is as follows:

$$\mathbb{E}_{\nu_{\mathrm{DE}}} f_{\mathrm{DE}}^{(1)}(\boldsymbol{\omega}, \mathbf{x})^2 f_{\mathrm{DE}}^{(2)}(\boldsymbol{\omega}, \mathbf{y})^2 = D^4 \det(\mathbf{I}_d - 8\mathbf{A})^{-1/2}$$
$$\times \exp\Bigg(2\mathbf{x}^\top \Big(\mathbf{C}^{(1)} + (\mathbf{B}^{(1)})^\top (\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(1)}\Big)\mathbf{x}$$
$$+2\mathbf{y}^\top \Big(\mathbf{C}^{(2)} + (\mathbf{B}^{(2)})^\top (\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(2)}\Big)\mathbf{y} + 4\mathbf{x}^\top (\mathbf{B}^{(1)})^\top (\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}^{(2)}\mathbf{y}\Bigg).$$

We conclude that the variance expression given in the theorem's statement is correct. □

A.2 Important lemma

Below, we prove an important lemma which is used in the subsequent proofs:

Lemma A.1. *Consider a function $f : (-\infty, \frac{1}{8})$ defined as*

$$f(A) = \log(1 - 4A) - \frac{1}{2}\log(1 - 8A) + \frac{\phi}{1 - 8A} \tag{14}$$

where $\phi \geq 0$. Then, the minimum of f on $(-\infty, \frac{1}{8})$ is achieved at

$$A^* = \frac{1}{16}\left(1 - 2\phi - \sqrt{(2\phi + 1)^2 + 8\phi}\right). \tag{15}$$

Proof. Set $\gamma = (1 - 8A)^{-1} \in (0, +\infty)$. Note that there is a one-to-one correspondence between $\gamma \in (0, +\infty)$ and $A \in (-\infty, \frac{1}{8})$. Hence, we can substitute $\gamma^{-1} = 1 - 8A$ and $1 - 4A = ((1 - 8A) + 1)/2 = (\gamma^{-1} + 1)/2 = \frac{1+\gamma}{2\gamma}$ in (14) and equivalently perform minimization with respect to γ:

$$\min_{\gamma \in (0, +\infty)} \quad h(\gamma) = \log\left(\frac{\gamma + 1}{2\gamma}\right) + \frac{1}{2}\log\gamma + \phi\gamma = \log(\gamma + 1) - \frac{1}{2}\log\gamma - \log 2 + \phi\gamma.$$

For $h(\cdot)$'s derivative, we have:

$$h'(\gamma) = \phi + \frac{1}{\gamma + 1} - \frac{1}{2\gamma} = \phi + \frac{\gamma - 1}{2\gamma(\gamma + 1)} \tag{16}$$
$$= \frac{2\phi\gamma(\gamma + 1) + \gamma - 1}{2\gamma(\gamma + 1)} = \frac{2\phi\gamma^2 + (2\phi + 1)\gamma - 1}{2\gamma(\gamma + 1)}. \tag{17}$$

Based on (16), we see that $h'(\gamma) \to -\infty$ as $\gamma \to 0$ and $h'(\gamma) > \phi \geq 0$ for all $\gamma > 1$. Hence, we conclude that $h(\cdot)$ is bounded from below on $(0, +\infty)$ and the global minimum γ^* on $(0, +\infty)$ exists and it is one of the points satisfying $h'(\gamma^*) = 0$. Hence, it's one of the positive roots of the polynomial in numerator of (17).

If $\phi = 0$, there is a single root $\gamma^* = 1$ of the polynomial in the numerator of (17), hence it is a global minimum of $h(\cdot)$. If $\phi > 0$, then there are two roots of the polynomial in the numerator of (17):

$$\gamma_-^* = \frac{-(2\phi + 1) - \sqrt{(2\phi + 1)^2 + 8\phi}}{4\phi},$$

$$\gamma_+^* = \frac{-(2\phi+1) + \sqrt{(2\phi+1)^2 + 8\phi}}{4\phi}. \tag{18}$$

Note that, if $\phi > 0$, then $2\phi + 1 > 0$ and $(2\phi+1)^2 + 8\phi \geq (2\phi+1)^2$. Hence, $\gamma_-^* < 0$ and $\gamma_+^* > 0$. We conclude that $\gamma^* = \gamma_+^*$ is the minimum of $h(\cdot)$ on $(0, +\infty)$. We multiply numerator and denominator of (18)'s right hand side by $(2\phi+1) + \sqrt{(2\phi+1)^2 + 8\phi} > 0$:

$$\gamma^* = \gamma_+^* = \frac{((2\phi+1)^2 + 8\phi) - (2\phi+1)^2}{4\phi\left((2\phi+1) + \sqrt{(2\phi+1)^2 + 8\phi}\right)} = \frac{2}{2\phi + 1 + \sqrt{(2\phi+1)^2 + 8\phi}}. \tag{19}$$

Note that the right hand side of (19) is equivalent to (18) when $\phi > 0$ but also holds for the case when $\phi = 0$ (i.e. when $\gamma^* = 1$). We conclude that $f(\cdot)$ is minimized at $\mathbf{A}^* = \frac{1}{8}(1 - (\gamma^*)^{-1})$ since $\gamma^* = (1 - 8A^*)^{-1}$. It's easy to see that (15) follows from (19) directly. □

A.3 Proof of Theorem 4.2

Proof. With $\mathbf{A} = A\mathbf{I}_d$, the conditions from Theorem 4.1 read as

$$8A < 1, \quad \frac{1}{1-4A}(\mathbf{B}^{(1)})^\top \mathbf{B}^{(2)} = \mathbf{I}_d, \quad \mathbf{C}^{(k)} = -\frac{1}{2(1-4A)}(\mathbf{B}^{(k)})^\top \mathbf{B}^{(k)}, \; D = (1-4A)^{d/4} \tag{20}$$

for $k \in \{1, 2\}$. And the variance expression (9) for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ transforms into

$$\mathrm{Var}_{\nu_{\mathrm{ADE}}} f_{\mathrm{ADE}}^{(1)}(\boldsymbol{\omega}, \mathbf{x}) f_{\mathrm{ADE}}^{(2)}(\boldsymbol{\omega}, \mathbf{y}) = D^4 (1-8A)^{-d/2} \exp\Bigg(2\mathbf{x}^\top \left(\mathbf{C}^{(1)} + \frac{1}{1-8A}(\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)}\right)\mathbf{x}$$
$$+2\mathbf{y}^\top \left(\mathbf{C}^{(2)} + \frac{1}{1-8A}(\mathbf{B}^{(2)})^\top \mathbf{B}^{(2)}\right)\mathbf{y} + \frac{4}{1-8A}\mathbf{x}^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(2)}\mathbf{y}\Bigg) - K^{(0)}(\mathbf{x}, \mathbf{y})^2.$$

We express $\mathbf{C}^{(k)}$ through $A, \mathbf{B}^{(k)}$ and D through A using (20) in the equation above:

$$\mathrm{Var}_{\nu_{\mathrm{ADE}}} f_{\mathrm{ADE}}^{(1)}(\boldsymbol{\omega}, \mathbf{x}) f_{\mathrm{ADE}}^{(2)}(\boldsymbol{\omega}, \mathbf{y}) = \left(\frac{1-4A}{\sqrt{1-8A}}\right)^d \exp\Bigg(\left(\frac{2}{1-8A} - \frac{1}{1-4A}\right)\mathbf{x}^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)}\mathbf{x}$$
$$+\left(\frac{2}{1-8A} - \frac{1}{1-4A}\right)\mathbf{y}^\top (\mathbf{B}^{(2)})^\top \mathbf{B}^{(2)}\mathbf{y} + \frac{4}{1-8A}\mathbf{x}^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(2)}\mathbf{y}\Bigg) - K^{(0)}(\mathbf{x}, \mathbf{y})^2.$$

Since $\frac{1}{1-4A}(\mathbf{B}^{(1)})^\top \mathbf{B}^{(2)}$ is a full-rank matrix $\mathbf{I}_d$ (20), both $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ are full-rank. Hence, we can express $\mathbf{B}^{(2)} = (1-4A)(\mathbf{B}^{(1)})^{-\top}$. Also, note that

$$\frac{2}{1-8A} - \frac{1}{1-4A} = \frac{2 - 8A - 1 + 8A}{(1-8A)(1-4A)} = (1-8A)^{-1}(1-4A)^{-1}.$$

We rewrite the expression for the variance using the identity above and the formula for $\mathbf{B}^{(2)}$:

$$\mathrm{Var}_{\nu_{\mathrm{ADE}}} f_{\mathrm{ADE}}^{(1)}(\boldsymbol{\omega}, \mathbf{x}) f_{\mathrm{ADE}}^{(2)}(\boldsymbol{\omega}, \mathbf{y}) = \left(\frac{1-4A}{\sqrt{1-8A}}\right)^d \exp\Bigg((1-8A)^{-1}(1-4A)^{-1}\mathbf{x}^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)}\mathbf{x}$$
$$+(1-8A)^{-1}(1-4A)\mathbf{y}^\top ((\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)})^{-1}\mathbf{y} + 4(1-8A)^{-1}(1-4A)\mathbf{x}^\top \mathbf{y}\Bigg) - K^{(0)}(\mathbf{x}, \mathbf{y})^2.$$

We use the expression above to rewrite (8) for $\langle \nu, f^{(1)}, f^{(2)} \rangle = \langle \nu_{\mathrm{ADE}}, f_{\mathrm{ADE}}^{(1)}, f_{\mathrm{ADE}}^{(2)} \rangle$ as follows:

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{ADE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{ADE}}) = L^{-2} \sum_{1 \leq i,j \leq L} \log(\mathrm{Var}_{\nu_{\mathrm{ADE}}} f_{\mathrm{ADE}}^{(1)}(\boldsymbol{\omega}, \mathbf{x}^{(i)}) f_{\mathrm{ADE}}^{(2)}(\boldsymbol{\omega}, \mathbf{y}^{(j)}) + K^{(0)}(\mathbf{x}^{(i)}, \mathbf{y}^{(j)}))$$
$$= d\log(1-4A) - \frac{d}{2}\log(1-8A) + (1-8A)^{-1}(1-4A)^{-1}L^{-1}\sum_{i=1}^{L}(\mathbf{x}^{(i)})^\top (\mathbf{B}^{(1)})^\top \mathbf{B}^{(1)}\mathbf{x}^{(i)}$$
$$+(1-8A)^{-1}(1-4A)L^{-1}\sum_{j=1}^{L}(\mathbf{y}^{(j)})^\top (\mathbf{B}^{(1)})^{-1}(\mathbf{B}^{(1)})^{-\top}\mathbf{y}^{(j)}$$

$$+4(1-8A)^{-1}(1-4A)L^{-2}\sum_{1\leq i,j\leq L}(\mathbf{x}^{(i)})^\top\mathbf{y}^{(j)}. \tag{21}$$

Denote $\mathbf{E}=(\mathbf{B}^{(1)})^\top\mathbf{B}^{(1)}$. Then (21) becomes:

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{ADE}};\mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{ADE}})=d\log(1-4A)-\frac{d}{2}\log(1-8A)$$
$$+(1-8A)^{-1}(1-4A)^{-1}L^{-1}\sum_{i=1}^{L}(\mathbf{x}^{(i)})^\top\mathbf{E}\mathbf{x}^{(i)}$$
$$+(1-8A)^{-1}(1-4A)L^{-1}\sum_{j=1}^{L}(\mathbf{y}^{(j)})^\top\mathbf{E}^{-1}\mathbf{y}^{(j)}+4(1-8A)^{-1}(1-4A)L^{-2}\sum_{1\leq i,j\leq L}(\mathbf{x}^{(i)})^\top\mathbf{y}^{(j)}. \tag{22}$$

We next prove the following lemma:

Lemma A.2. *Let* $\mathbf{B}^{(1)*}=\sqrt{1-4A}\boldsymbol{\Sigma}^{1/2}\mathbf{U}^\top(\boldsymbol{\Lambda}^{(1)})^{-1/2}(\mathbf{Q}^{(1)})^\top$. *When* A *(*$8A<1$*) is fixed,* $\mathbf{E}=\mathbf{E}^*=(\mathbf{B}^{(1)*})^\top\mathbf{B}^{(1)*}$ *minimizes the right hand side of (22) with respect to* $\mathbf{E}$*.*

Proof. We have:

$$L^{-1}\sum_{i=1}^{L}(\mathbf{x}^{(i)})^\top\mathbf{E}\mathbf{x}^{(i)}=L^{-1}\sum_{i=1}^{L}\mathrm{Trace}((\mathbf{x}^{(i)})^\top\mathbf{E}\mathbf{x}^{(i)})=L^{-1}\sum_{i=1}^{L}\mathrm{Trace}(\mathbf{E}\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top)$$
$$=\mathrm{Trace}\left(\mathbf{E}\left(L^{-1}\sum_{i=1}^{L}\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top\right)\right)=\mathrm{Trace}(\mathbf{E}\mathbf{M}^{(1)})$$

where we use the cyclic property of trace $\mathrm{Trace}(\cdot)$ and linearity of trace. Analogously, we obtain $L^{-1}\sum_{j=1}^{L}(\mathbf{y}^{(j)})^\top\mathbf{E}^{-1}\mathbf{y}^{(j)}=\mathrm{Trace}(\mathbf{E}^{-1}\mathbf{M}^{(2)})$. Assuming that A is fixed, optimization of (22) with respect to $\mathbf{E}$ reduces to the following minimization problem:

$$\min_{\mathbf{E}\in\mathbb{S}_d,\mathbf{E}\succ 0}\quad\mathcal{F}(\mathbf{E})=\beta_1\mathrm{Trace}(\mathbf{E}\mathbf{M}^{(1)})+\beta_2\mathrm{Trace}(\mathbf{E}^{-1}\mathbf{M}^{(2)}) \tag{23}$$

where $\beta_1=(1-8A)^{-1}(1-4A)^{-1}$, $\beta_2=(1-8A)^{-1}(1-4A)$ and the constraint $\mathbf{E}\in\mathbb{S}_d,\mathbf{E}\succ 0$ follows from the fact that $\mathbf{E}=(\mathbf{B}^{(1)})^\top\mathbf{B}^{(1)}$ and $\mathbf{E}$ is invertible. We have $1-8A>0$ and $1-4A=(1-8A)/2+1/2>0$. Hence, $\beta_1,\beta_2>0$. For any $\mathbf{E}\succ 0$ and any $\boldsymbol{\Delta}\in\mathbb{S}_d$ there is $t\in\mathbb{R}$ small enough such that $\mathbf{E}+t\mathbf{B}$ is invertible and the following Neumann series is convergent:

$$(\mathbf{E}+t\boldsymbol{\Delta})^{-1}=\mathbf{E}^{-1}(\mathbf{I}_d+t\boldsymbol{\Delta}\mathbf{E}^{-1})^{-1}=\sum_{l=0}^{\infty}(-t)^l\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l$$

We further deduce:

$$\mathrm{Trace}((\mathbf{E}+t\boldsymbol{\Delta})^{-1}\mathbf{M}^{(2)})=\mathrm{Trace}\left(\left(\sum_{l=0}^{\infty}(-t)^l\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\right)\mathbf{M}^{(2)}\right)=$$
$$=\sum_{l=0}^{\infty}(-t)^l\mathrm{Trace}\left(\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\mathbf{M}^{(2)}\right)$$

and, therefore,

$$\mathcal{F}(\mathbf{E}+t\boldsymbol{\Delta})=\beta_1\mathrm{Trace}((\mathbf{E}+t\boldsymbol{\Delta})\mathbf{M}^{(1)})+\beta_2\sum_{l=0}^{\infty}(-t)^l\mathrm{Trace}\left(\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\mathbf{M}^{(2)}\right)$$
$$=\beta_1\mathrm{Trace}(\mathbf{E}\mathbf{M}^{(1)})+t\beta_1\mathrm{Trace}(\boldsymbol{\Delta}\mathbf{M}^{(1)})+\beta_2\sum_{l=0}^{\infty}(-t)^l\mathrm{Trace}\left(\mathbf{E}^{-1}(\boldsymbol{\Delta}\mathbf{E}^{-1})^l\mathbf{M}^{(2)}\right). \tag{24}$$

Further, we have:

$$\frac{\partial}{\partial t}\mathcal{F}(\mathbf{E}+t\mathbf{\Delta}) = \beta_1 \text{Trace}(\mathbf{\Delta}\mathbf{M}^{(1)}) + \beta_2 \sum_{l=1}^{\infty}(-1)^l l t^{l-1} \text{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right),$$

$$\frac{\partial^2}{(\partial t)^2}\mathcal{F}(\mathbf{E}+t\mathbf{\Delta}) = \beta_2 \sum_{l=2}^{\infty}(-1)^l l(l-1) t^{l-2} \text{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right),$$

$$\frac{\partial^2}{(\partial t)^2}\mathcal{F}(\mathbf{E}+t\mathbf{\Delta})\Big|_{t=0} = 2\beta_2 \text{Trace}\left(\mathbf{E}^{-1}\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{M}^{(2)}\right). \tag{25}$$

We replace $\mathbf{M}^{(2)} = \mathbf{Q}^{(2)}(\mathbf{\Lambda}^{(2)})^{1/2}(\mathbf{\Lambda}^{(2)})^{1/2}(\mathbf{Q}^{(2)})^\top$ and apply the cyclic property of trace in (25):

$$\frac{\partial^2}{(\partial t)^2}\mathcal{F}(\mathbf{E}+t\mathbf{\Delta})\Big|_{t=0} = 2\beta_2 \text{Trace}\left((\mathbf{\Lambda}^{(2)})^{1/2}(\mathbf{Q}^{(2)})^\top \mathbf{E}^{-1}\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{Q}^{(2)}(\mathbf{\Lambda}^{(2)})^{1/2}\right)$$
$$= 2\beta_2 \text{Trace}\left(\mathbf{T}\mathbf{E}^{-1}\mathbf{T}^\top\right)$$

where $\mathbf{T} = (\mathbf{\Lambda}^{(2)})^{1/2}(\mathbf{Q}^{(2)})^\top \mathbf{E}^{-1}\mathbf{\Delta}$. Since $\mathbf{E}$ is positive definite, $\mathbf{E}^{-1}$ is also positive definite and $\mathbf{T}\mathbf{E}^{-1}\mathbf{T}^\top$ is at least positive semidefinite. Hence, $\text{Trace}\left(\mathbf{T}\mathbf{E}^{-1}\mathbf{T}^\top\right) \geq 0$ and also $\frac{\partial^2}{(\partial t)^2}\mathcal{F}(\mathbf{E}+t\mathbf{\Delta})|_{t=0} \geq 0$. We conclude that $\mathcal{F}(\mathbf{E})$ is a convex function on $\{\mathbf{E} \in \mathbb{S}_d \,|\, \mathbf{E} \succ 0\}$. Since $\{\mathbf{E} \in \mathbb{S}_d \,|\, \mathbf{E} \succ 0\}$ is an open set, (every) global minimum $\mathbf{E}$ of (23) satisfies two conditions

$$1)\ \mathbf{E} \succ 0, \quad \text{and} \quad 2)\ \nabla\mathcal{F}(\mathbf{E}) = \mathbf{0}_{d\times d} \tag{26}$$

Set $t = 1$ and assume that $\mathbf{\Delta} \in \mathbb{S}_d$ is small enough by norm so that $\mathbf{E}+\mathbf{\Delta}$ is invertible and the Neumann series for $(\mathbf{I}_d + \mathbf{\Delta}\mathbf{E}^{-1})^{-1}$ is convergent. Then, (24) holds for $t = 1$:

$$\mathcal{F}(\mathbf{E}+\mathbf{\Delta}) = \beta_1 \text{Trace}(\mathbf{E}\mathbf{M}^{(1)}) + \beta_1 \text{Trace}(\mathbf{\Delta}\mathbf{M}^{(1)}) + \beta_2 \sum_{l=0}^{\infty}(-1)^l \text{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right)$$
$$= \mathcal{F}(\mathbf{E}) + \beta_1 \text{Trace}(\mathbf{\Delta}\mathbf{M}^{(1)}) + \beta_2 \sum_{l=1}^{\infty}(-1)^l \text{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right)$$
$$= \mathcal{F}(\mathbf{E}) + \beta_1 \text{Trace}(\mathbf{\Delta}\mathbf{M}^{(1)}) - \beta_2 \text{Trace}\left(\mathbf{E}^{-1}\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{M}^{(2)}\right)$$
$$+\beta_2 \sum_{l=2}^{\infty}(-1)^l \text{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right).$$

Clearly, $\beta_2 \sum_{l=2}^{\infty}(-1)^l \text{Trace}\left(\mathbf{E}^{-1}(\mathbf{\Delta}\mathbf{E}^{-1})^l \mathbf{M}^{(2)}\right) = o(\|\mathbf{\Delta}\|)$ where $\|\cdot\|$ is an L_2-norm. Also, using the cyclic property of trace, we get:

$$\text{Trace}\left(\mathbf{E}^{-1}\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{M}^{(2)}\right) = \text{Trace}\left(\mathbf{\Delta}\mathbf{E}^{-1}\mathbf{M}^{(2)}\mathbf{E}^{-1}\right).$$

Therefore, we have:

$$\mathcal{F}(\mathbf{E}+\mathbf{\Delta}) = \text{Trace}\left(\mathbf{\Delta}\left(\beta_1\mathbf{M}^{(1)} - \beta_2\mathbf{E}^{-1}\mathbf{M}^{(2)}\mathbf{E}^{-1}\right)\right) + o(\|\mathbf{\Delta}\|). \tag{27}$$

Since $\mathbf{\Delta}, \mathbf{E}^{-1}, \mathbf{M}^{(1)}, \mathbf{M}^{(2)} \in \mathbb{S}_d$, from (27) it follows that

$$\nabla\mathcal{F}(\mathbf{E}) = \beta_1\mathbf{M}^{(1)} - \beta_2\mathbf{E}^{-1}\mathbf{M}^{(2)}\mathbf{E}^{-1}. \tag{28}$$

Let $\mathbf{E}^* = (\mathbf{B}^{(1)*})^\top \mathbf{B}^{(1)*} \succeq 0$. Note that

$$\sqrt[4]{\frac{\beta_2}{\beta_1}} = \sqrt[4]{\frac{(1-8A)^{-1}(1-4A)}{(1-8A)^{-1}(1-4A)^{-1}}} = \sqrt{1-4A}.$$

Since $\sqrt{\beta_2/\beta_1} \neq 0$, $\mathbf{\Sigma}$, $\mathbf{U}$, $\mathbf{\Lambda}^{-1/2}$, $\mathbf{Q}^{(1)}$ are full-rank, $\mathbf{E}^*$ is also full-rank, therefore $\mathbf{E}^* \succ 0$ and it satisfies condition 1 from (26). Observe that

$$\mathbf{E}^*\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{1/2} = \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{-1/2}(\mathbf{Q}^{(1)})^\top\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{1/2}$$

$$
\begin{aligned}
&= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top \\
&= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)\mathbf{V}\mathbf{U}^\top \\
&= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}((\mathbf{\Lambda}^{(1)})^{\frac{1}{2}}(\mathbf{Q}^{(1)})^\top\mathbf{Q}^{(2)}(\mathbf{\Lambda}^{(2)})^{\frac{1}{2}})\mathbf{V}\mathbf{U}^\top \\
&= \sqrt{\beta_2/\beta_1}\mathbf{Q}^{(2)}(\mathbf{\Lambda}^{(2)})^{1/2}\mathbf{V}\mathbf{U}^\top
\end{aligned}
$$

where we use definitions of $\mathbf{E}^*$, $\mathbf{U}$, $\mathbf{\Sigma}$, $\mathbf{V}$ and orthogonality of $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \mathbf{U}, \mathbf{V}$. Hence, we deduce that

$$
\begin{aligned}
\beta_1\mathbf{E}^*\mathbf{M}^{(1)}\mathbf{E}^* &= \beta_1\mathbf{E}^*\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{1/2}\left((\mathbf{\Lambda}^{(1)})^{1/2}(\mathbf{Q}^{(1)})^\top\mathbf{E}^*\right) && (29) \\
&= \beta_1\frac{\beta_2}{\beta_1}\mathbf{Q}^{(2)}(\mathbf{\Lambda}^{(2)})^{1/2}\left((\mathbf{\Lambda}^{(2)})^{1/2}(\mathbf{Q}^{(2)})^\top\right) \\
&= \beta_2\mathbf{M}^{(2)} && (30)
\end{aligned}
$$

by the definition of $\mathbf{Q}^{(1)}$, $\mathbf{\Lambda}^{(1)}$, $\mathbf{Q}^{(2)}$, $\mathbf{\Lambda}^{(2)}$ and due to orthogonality of $\mathbf{V}$, $\mathbf{U}$. By left- and right-multiplication of (30) by $(\mathbf{E}^*)^{-1}$ we deduce that

$$
\beta_1\mathbf{M}^{(1)} = \beta_2(\mathbf{E}^*)^{-1}\mathbf{M}^{(2)}(\mathbf{E}^*)^{-1}
$$

or, in other words, $\nabla\mathcal{F}(\mathbf{E}^*) = \mathbf{0}_{d\times d}$ and the condition 2 from (26) is also satisfied. We conclude that the global minimum of (23) is achieved at $\mathbf{E}^*$. □

According to Lemma A.2, $\mathbf{B}^{(1)} = \mathbf{B}^{(1)*}$ is a global minimum of (21)'s right hand side when A is fixed. Indeed, if there is $\mathbf{B}^{(1)}$ which leads to a smaller value of (21), $\mathbf{E} = (\mathbf{B}^{(1)})^\top\mathbf{B}^{(1)}$ would lead to a smaller value of (22)'s right hand side. Also, this $\mathbf{E}$ is positive definite by definition (note that $\mathbf{B}^{(1)}$ is nonsingular), leading to contradiction with Lemma A.2.

Substituting $\mathbf{E}^*$ instead of $\mathbf{E}$ in (22) corresponds to the minimum value of $\overline{\mathcal{L}}(\boldsymbol{\theta}_{\text{AGE}}; \alpha, \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\text{AGE}})$ for a fixed A. Our next step is to minimize this expression with respect to A. Denote $\mathbf{F} = \mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{-1/2}(\mathbf{Q}^{(1)})^\top$. Then $\mathbf{E}^* = (1-4A)\mathbf{F}$ where $\mathbf{F}$ doesn't depend on A. We substitute $\mathbf{E}^*$ into (22) and get:

$$
\begin{aligned}
&d\log(1-4A) - \frac{d}{2}\log(1-8A) + (1-8A)^{-1}(1-4A)^{-1}\text{Trace}((1-4A)\mathbf{F}\mathbf{M}^{(1)}) \\
&\quad +(1-8A)^{-1}(1-4A)\text{Trace}((1-4A)^{-1}\mathbf{F}^{-1}\mathbf{M}^{(2)}) \\
&\quad +4(1-8A)^{-1}(1-4A)L^{-2}\sum_{1\le i,j\le L}(\mathbf{x}^{(i)})^\top\mathbf{y}^{(j)} \\
= d\log(1-4A) - \frac{d}{2}\log(1-8A) &+ (1-8A)^{-1}\text{Trace}(\mathbf{F}\mathbf{M}^{(1)}) + (1-8A)^{-1}\text{Trace}(\mathbf{F}^{-1}\mathbf{M}^{(2)}) \\
&+2\left(1+(1-8A)^{-1}\right)d\mu^{(3)} && (31)
\end{aligned}
$$

where we also replace

$$
L^{-2}\sum_{1\le i,j\le L}(\mathbf{x}^{(i)})^\top\mathbf{y}^{(j)} = L^{-2}\left(\sum_{i=1}^{L}\mathbf{x}^{(i)}\right)^\top\left(\sum_{j=1}^{L}\mathbf{y}^{(j)}\right) = d\mu^{(3)}
$$

and

$$
(1-8A)^{-1}(1-4A) = \frac{(1-8A)+1}{2(1-8A)} = \frac{1}{2}\left(1+(1-8A)^{-1}\right)
$$

Based on (30) and since $\mathbf{F} = \sqrt{\beta_1/\beta_2}\mathbf{E}^*$, we conclude that $\mathbf{F}\mathbf{M}^{(1)}\mathbf{F} = \mathbf{M}^{(2)}$, or $\mathbf{M}^{(1)}\mathbf{F} = \mathbf{F}^{-1}\mathbf{M}^{(2)}$. Using the cyclic property of trace, we get:

$$
\text{Trace}(\mathbf{F}\mathbf{M}^{(1)}) = \text{Trace}(\mathbf{M}^{(1)}\mathbf{F}) = \text{Trace}(\mathbf{F}^{-1}\mathbf{M}^{(2)}).
$$

By the definition of $\mathbf{F}$, $\mathbf{\Lambda}^{(1)}, \mathbf{Q}^{(1)}$ and using the cyclic property and orthogonality of $\mathbf{Q}^{(1)}, \mathbf{U}$, we have:

$$
\text{Trace}(\mathbf{F}\mathbf{M}^{(1)}) = \text{Trace}\left(\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{Q}^{(1)})^\top\left(\mathbf{Q}^{(1)}\mathbf{\Lambda}^{(1)}(\mathbf{Q}^{(1)})^\top\right)\right)
$$

$$= \text{Trace}\left(\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{1/2}(\mathbf{Q}^{(1)})^\top\right)$$
$$= \text{Trace}\left(\mathbf{\Sigma}\mathbf{U}^\top(\mathbf{\Lambda}^{(1)})^{1/2}(\mathbf{Q}^{(1)})^\top\mathbf{Q}^{(1)}(\mathbf{\Lambda}^{(1)})^{-1/2}\mathbf{U}\right)$$
$$= \text{Trace}(\mathbf{\Sigma}) = \sum_{l=1}^{d}\mathbf{\Sigma}_{l,l}.$$

Hence, (31) finally becomes:

$$d\log(1-4A) - \frac{d}{2}\log(1-8A) + 2(1-8A)^{-1}\sum_{l=1}^{d}\mathbf{\Sigma}_{l,l} + 2\left(1+(1-8A)^{-1}\right)d\mu^{(3)}$$
$$= d\left(\log(1-4A) - \frac{1}{2}\log(1-8A) + 2(1-8A)^{-1}\left(d^{-1}\sum_{l=1}^{d}\mathbf{\Sigma}_{l,l} + \mu^{(3)}\right) + 2\mu^{(3)}\right). \quad (32)$$

Next, we use Lemma A.1 ($\phi = d^{-1}\sum_{l=1}^{d}\mathbf{\Sigma}_{l,l} + \mu^{(3)} \geq 0$) for deriving expression for A which minimizes (32). This expression coincides with the one in Theorem's statement. The expressions for $\mathbf{B}^{(2)}, \mathbf{C}^{(1)}, \mathbf{C}^{(2)}$ follow directly from (20), optimal $\mathbf{B}^{(1)} = \mathbf{B}^{(1)*}$ and A. (10) follows from (32). The proof is concluded. □

A.4 Proof of Theorem 4.3

Proof. With $\mathbf{B}^{(1)} = \mathbf{B}^{(2)} = \mathbf{B}$ and $\mathbf{C}^{(1)} = \mathbf{C}^{(2)} = \mathbf{C}$, the conditions from Theorem 4.1 read as

$$8\mathbf{A} \prec \mathbf{I}_d, \quad \mathbf{B}^\top(\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B} = \mathbf{I}_d, \quad \mathbf{C} = -\frac{1}{2}\mathbf{B}^\top(\mathbf{I}_d - 4\mathbf{A})^{-1}\mathbf{B} = -\frac{1}{2}\mathbf{I}_d, \quad D = \det(\mathbf{I}_d - 4\mathbf{A})^{1/4}. \quad (33)$$

Denote $\mathbf{Q} = (\mathbf{I}_d - 4\mathbf{A})^{-1/2}\mathbf{B} \in \mathbb{R}^{d\times d}$. Then, according to (33), $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_d$, that is $\mathbf{Q} \in \mathbb{O}_d$. We rewrite (9) using (33) and then substitute $\mathbf{B} = (\mathbf{I}_d - 4\mathbf{A})^{1/2}\mathbf{Q}$:

$$\text{Var}_{\nu_{\text{SDE}}} f^{(1)}_{\text{SDE}}(\boldsymbol{\omega}, \mathbf{x}) f^{(2)}_{\text{SDE}}(\boldsymbol{\omega}, \mathbf{y}) = \det(\mathbf{I}_d - 4\mathbf{A})\det(\mathbf{I}_d - 8\mathbf{A})^{-1/2}\exp\Big(-\|\mathbf{x}\|^2$$
$$+2\mathbf{x}^\top\mathbf{B}^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}\mathbf{x} - \|\mathbf{y}\|^2 + 2\mathbf{y}^\top\mathbf{B}^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}\mathbf{y} + 4\mathbf{x}^\top\mathbf{B}^\top(\mathbf{I}_d - 8\mathbf{A})^{-1}\mathbf{B}\mathbf{y}\Big)$$
$$-K^{(0)}(\mathbf{x}, \mathbf{y})^2 = \det(\mathbf{I}_d - 4\mathbf{A})^{1/4}\det(\mathbf{I}_d - 8\mathbf{A})^{-1/2}\exp\Big(-\|\mathbf{x}\|^2 - 2\mathbf{x}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{x} - \|\mathbf{y}\|^2$$
$$-2\mathbf{y}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{y} - 4\mathbf{x}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{y}\Big) - K^{(0)}(\mathbf{x}, \mathbf{y})^2 \quad (34)$$

where we denote:

$$\mathbf{E} = -(\mathbf{I}_d - 4\mathbf{A})^{1/2}(\mathbf{I}_d - 8\mathbf{A})^{-1}(\mathbf{I}_d - 4\mathbf{A})^{1/2} = -(\mathbf{I}_d - 4\mathbf{A})(\mathbf{I}_d - 8\mathbf{A})^{-1}$$
$$= -\frac{1}{2}\left((\mathbf{I}_d - 8\mathbf{A}) + \mathbf{I}_d\right)(\mathbf{I}_d - 8\mathbf{A})^{-1} = -\frac{1}{2}\mathbf{I}_d - \frac{1}{2}(\mathbf{I}_d - 8\mathbf{A})^{-1} \quad (35)$$

which is in $\mathbb{D}_d$ since $\mathbf{A} \in \mathbb{D}_d$. Next, we observe:

$$2\mathbf{x}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{x} + 2\mathbf{y}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{y} + 4\mathbf{x}^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{y} = 2(\mathbf{x}+\mathbf{y})^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}+\mathbf{y})$$

We plug this into (34) and use the resulting expression to rewrite (8) for $\langle \nu, f^{(1)}, f^{(2)}\rangle = \langle \nu_{\text{SDE}}, f^{(1)}_{\text{SDE}}, f^{(2)}_{\text{SDE}}\rangle$ as follows:

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\text{SDE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\text{SDE}}) = L^{-2}\sum_{1\leq i,j\leq L}\log(\text{Var}_{\nu_{\text{SDE}}} f^{(1)}_{\text{SDE}}(\boldsymbol{\omega}, \mathbf{x}^{(i)}) f^{(2)}_{\text{SDE}}(\boldsymbol{\omega}, \mathbf{y}^{(j)}) + K^{(0)}(\mathbf{x}^{(i)}, \mathbf{y}^{(j)}))$$
$$= \log\det(\mathbf{I}_d - 4\mathbf{A}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2$$
$$-2L^{-2}\sum_{1\leq i,j\leq L}(\mathbf{x}^{(i)} + \mathbf{y}^{(j)})^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}^{(i)} + \mathbf{y}^{(j)}). \quad (36)$$

Using linearity and cyclic property of trace, we deduce that

$$L^{-2}\sum_{1\leq i,j\leq L}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})=$$
$$=L^{-2}\sum_{1\leq i,j\leq L}\mathrm{Trace}\left((\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})\right)$$
$$=L^{-2}\sum_{1\leq i,j\leq L}\mathrm{Trace}\left(\mathbf{Q}^\top\mathbf{E}\mathbf{Q}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\right)$$
$$=\mathrm{Trace}\left(\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\left(L^{-2}\sum_{1\leq i,j\leq L}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top\right)\right)$$

Observe that

$$L^{-2}\sum_{1\leq i,j\leq L}(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})(\mathbf{x}^{(i)}+\mathbf{y}^{(j)})^\top=$$
$$=L^{-2}\sum_{1\leq i,j\leq L}\left(\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top+\mathbf{x}^{(i)}(\mathbf{y}^{(j)})^\top+\mathbf{y}^{(j)}(\mathbf{x}^{(i)})^\top+\mathbf{y}^{(j)}(\mathbf{x}^{(j)})^\top\right)$$
$$=L^{-1}\sum_{i=1}^{L}\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\top+\left(L^{-1}\sum_{i=1}^{L}\mathbf{x}^{(i)}\right)\left(L^{-1}\sum_{j=1}^{L}\mathbf{y}^{(j)}\right)^\top+\left(L^{-1}\sum_{j=1}^{L}\mathbf{y}^{(j)}\right)\left(L^{-1}\sum_{i=1}^{L}\mathbf{x}^{(i)}\right)^\top$$
$$+L^{-1}\sum_{j=1}^{L}\mathbf{y}^{(j)}(\mathbf{y}^{(j)})^\top=\mathbf{M}^{(1)}+\boldsymbol{\mu}^{(4)}(\boldsymbol{\mu}^{(5)})^\top+\boldsymbol{\mu}^{(5)}(\boldsymbol{\mu}^{(4)})^\top+\mathbf{M}^{(2)}.$$

Denote $\mathbf{N}=\mathbf{M}^{(1)}+\boldsymbol{\mu}^{(4)}(\boldsymbol{\mu}^{(5)})^\top+\boldsymbol{\mu}^{(5)}(\boldsymbol{\mu}^{(4)})^\top+\mathbf{M}^{(2)}$. We conclude that

$$\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{SDE}};\mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{SDE}})=\log\det(\mathbf{I}_d-4\mathbf{A})-\frac{1}{2}\log\det(\mathbf{I}_d-8\mathbf{A})-L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2$$
$$-L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2-2\mathrm{Trace}\left(\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{N}\right). \tag{37}$$

With $\mathbf{A}$ fixed, we minimize the right hand side of (37) with respect to $\mathbf{Q}$ which is equivalent to minimizing $\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{SDE}};\mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{SDE}})$ with respect to $\mathbf{B}$ with fixed $\mathbf{A}$, since there is a one-to-one correspondence between $\mathbf{B}$ and $\mathbf{Q}$. This is equivalent to maximizing, again using the cyclic property of trace,

$$\mathrm{Trace}\left(\mathbf{Q}^\top\mathbf{E}\mathbf{Q}\mathbf{N}\right)=\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}\mathbf{N}\mathbf{Q}^\top\right) \tag{38}$$

with respect to $\mathbf{Q}$. We prove the following lemma first:

Lemma A.3. *Suppose that diagonal entries of $\mathbf{E}$ are all distinct, and the same holds for $\boldsymbol{\Lambda}^{(3)}$. Let $\boldsymbol{\Pi}\in\{0,1\}^{d\times d}$ be a permutation matrix sorting diagonal entries of $\mathbf{E}$ (i.e. by applying $\boldsymbol{\Pi}\mathbf{E}\boldsymbol{\Pi}^\top$) in a descending order corresponding to a permutation $\boldsymbol{\pi}\in\mathbb{N}^d$. Set $\mathbf{Q}^*=\boldsymbol{\Pi}^\top(\mathbf{Q}^{(3)})^\top\in\mathbb{O}_d$. Then we have:*

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^*\mathbf{N}(\mathbf{Q}^*)^\top\right)=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\boldsymbol{\Lambda}^{(3)}_{l,l} \tag{39}$$
$$=\sup_{\mathbf{Q}\in\mathbb{O}_d}\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}\mathbf{N}\mathbf{Q}^\top\right) \tag{40}$$

Proof. First of all, we have:

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^*\mathbf{N}(\mathbf{Q}^*)^\top\right)=\mathrm{Trace}\left(\mathbf{E}\boldsymbol{\Pi}^\top(\mathbf{Q}^{(3)})^\top\mathbf{N}\mathbf{Q}^{(3)}\boldsymbol{\Pi}\right)=\mathrm{Trace}\left(\mathbf{E}\boldsymbol{\Pi}^\top\boldsymbol{\Lambda}^{(3)}\boldsymbol{\Pi}\right) \tag{41}$$

$$= \text{Trace}\left(\mathbf{\Pi E \Pi}^\top \mathbf{\Lambda}^{(3)}\right) = \sum_{l=1}^{d} \mathbf{E}_{\boldsymbol{\pi}_l, \boldsymbol{\pi}_l} \mathbf{\Lambda}^{(3)}_{l,l}, \tag{42}$$

i.e. (39) is satisfied.

Optimization for finding $\sup_{\mathbf{Q} \in \mathbb{O}_d} \text{Trace}\left(\mathbf{EQNQ}^\top\right)$ is a well-studied problem [8]. By the definition, $\mathbf{\Lambda}^{(3)}$ has eigenvalues of $\mathbf{N}$ on the main diagonal and $\mathbf{E} \in \mathbb{D}_d$ hence it contains its eigenvalues on its main diagonal. Then, as proven in [8], $\mathbf{Q}^*$ is indeed a global maximum of this problem in the case of distinct eigenvalues for $\mathbf{E}$ and $\mathbf{N}$. That is, (40) is proven. □

Next, we prove a generalization of Lemma A.3 when diagonal entries of $\mathbf{E}$ and $\mathbf{\Lambda}^{(3)}$ are not necessarily distinct:

Lemma A.4. *Let $\mathbf{\Pi} \in \{0,1\}^{d \times d}$ be a permutation matrix sorting diagonal entries of $\mathbf{E}$ (i.e. by applying $\mathbf{\Pi E \Pi}^\top$) in **any non-ascending** order corresponding to a permutation $\boldsymbol{\pi} \in \mathbb{N}^d$. Set $\mathbf{Q}^* = \mathbf{\Pi}^\top (\mathbf{Q}^{(3)})^\top \in \mathbb{O}_d$. Then we have:*

$$\text{Trace}\left(\mathbf{EQ}^*\mathbf{N}(\mathbf{Q}^*)^\top\right) = \sum_{l=1}^{d} \mathbf{E}_{\boldsymbol{\pi}_l, \boldsymbol{\pi}_l} \mathbf{\Lambda}^{(3)}_{l,l} \tag{43}$$

$$= \sup_{\mathbf{Q} \in \mathbb{O}_d} \text{Trace}\left(\mathbf{EQNQ}^\top\right) \tag{44}$$

Proof. In the same way as (41-42), we show that $\text{Trace}\left(\mathbf{EQ}^*\mathbf{N}(\mathbf{Q}^*)^\top\right) = \sum_{l=1}^{d} \mathbf{E}_{\boldsymbol{\pi}_l, \boldsymbol{\pi}_l} \mathbf{\Lambda}^{(3)}_{l,l}$, i.e. (43) is satisfied. Next we prove that for any $\mathbf{Q} \in \mathbb{O}_d$,

$$\text{Trace}\left(\mathbf{EQNQ}^\top\right) \leq \sum_{l=1}^{d} \mathbf{E}_{\boldsymbol{\pi}_l, \boldsymbol{\pi}_l} \mathbf{\Lambda}^{(3)}_{l,l}. \tag{45}$$

which would imply (44).

Our proof is by contradiction. First of all, we can assume that $\mathbf{E}, \mathbf{\Lambda}^{(3)}$ are nonzero matrices since otherwise we have (44) trivially. Since $\text{Trace}\left(\mathbf{EQNQ}^\top\right)$ is a continuous function of $\mathbf{Q}$ and $\mathbb{O}_d$ is compact, $\sup_{\mathbf{Q} \in \mathbb{O}_d} \text{Trace}\left(\mathbf{EQNQ}^\top\right)$ is finite. Suppose that there is $\delta > 0$ such that

$$\delta = \sup_{\mathbf{Q} \in \mathbb{O}_d} \text{Trace}\left(\mathbf{EQNQ}^\top\right) - \sum_{l=1}^{d} \mathbf{E}_{\boldsymbol{\pi}_l, \boldsymbol{\pi}_l} \mathbf{\Lambda}^{(3)}_{l,l}. \tag{46}$$

Let $\widetilde{\mathbf{E}}, \widetilde{\mathbf{\Lambda}}^{(3)} \in \mathbb{D}_d$ be matrices with all distinct values on the diagonal such that

$$\|\widetilde{\mathbf{E}} - \mathbf{E}\|_{\text{F}} \leq \min\left(\|\mathbf{E}\|_{\text{F}}, \frac{\delta}{12\|\mathbf{\Lambda}^{(3)}\|_{\text{F}}}\right), \quad \|\widetilde{\mathbf{\Lambda}}^{(3)} - \mathbf{\Lambda}^{(3)}\|_{\text{F}} \leq \frac{\delta}{12\|\mathbf{E}\|_{\text{F}}} \tag{47}$$

where $\|\cdot\|_{\text{F}}$ denotes Frobenius norm and $\|\mathbf{E}\|_{\text{F}}, \|\mathbf{\Lambda}^{(3)}\|_{\text{F}} \neq 0$ since these are nonzero matrices. Further, we assume that diagonal entries of $\widetilde{\mathbf{\Lambda}}^{(3)}$ are sorted in a descending order and, in addition to $\mathbf{\Lambda}^{(3)}$, $\boldsymbol{\pi}$ also sorts entries of $\widetilde{\mathbf{E}}$ in a non-ascending (descending) order. Clearly, such $\widetilde{\mathbf{E}}, \widetilde{\mathbf{\Lambda}}^{(3)}$ can be obtained by small perturbations of $\mathbf{E}, \mathbf{\Lambda}^{(3)}$. Also, denote $\widetilde{\mathbf{N}} = \mathbf{Q}^{(3)}\widetilde{\mathbf{\Lambda}}^{(3)}(\mathbf{Q}^{(3)})^\top$. Since $\mathbb{O}_d$ is a compact closed set and $\text{Trace}\left(\mathbf{EQNQ}^\top\right)$ is a continuous function of $\mathbf{Q}$, there exists $\mathbf{Q}^{**} \in \mathbb{O}_d$ such that

$$\text{Trace}\left(\mathbf{EQ}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) = \sup_{\mathbf{Q} \in \mathbb{O}_d} \text{Trace}\left(\mathbf{EQNQ}^\top\right). \tag{48}$$

By the definition of $\widetilde{\mathbf{E}}, \widetilde{\mathbf{\Lambda}}^{(3)}, \widetilde{\mathbf{N}}$, we have:

$$\text{Trace}\left(\mathbf{EQ}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) - \text{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)$$

$$= \left(\text{Trace}\left(\mathbf{EQ}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right) - \text{Trace}\left(\mathbf{EQ}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)\right) + \left(\text{Trace}\left(\mathbf{EQ}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)\right.$$

$$-\mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)$$
$$=\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\left(\mathbf{N}-\widetilde{\mathbf{N}}\right)(\mathbf{Q}^{**})^\top\right)+\mathrm{Trace}\left(\left(\mathbf{E}-\widetilde{\mathbf{E}}\right)\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right).$$

Next, we apply Cauchy-Schwarz inequality to both terms:

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\left(\mathbf{N}-\widetilde{\mathbf{N}}\right)(\mathbf{Q}^{**})^\top\right)\leq\|(\mathbf{Q}^{**})^\top\mathbf{E}\|_{\mathrm{F}}\|(\mathbf{N}-\widetilde{\mathbf{N}})(\mathbf{Q}^{**})^\top\|_{\mathrm{F}}=\|\mathbf{E}\|_{\mathrm{F}}\|\mathbf{N}-\widetilde{\mathbf{N}}\|_{\mathrm{F}},$$
$$\mathrm{Trace}\left(\left(\mathbf{E}-\widetilde{\mathbf{E}}\right)\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)\leq\|(\mathbf{Q}^{**})^\top(\mathbf{E}-\widetilde{\mathbf{E}})\|_{\mathrm{F}}\|\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\|_{\mathrm{F}}=\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{N}}\|_{\mathrm{F}}$$

where we use invariance of the Frobenius norm under multiplications by orthogonal matrices. Using this invariance again, we deduce that

$$\|\mathbf{N}-\widetilde{\mathbf{N}}\|_{\mathrm{F}}=\|\mathbf{Q}^{(3)}(\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)})(\mathbf{Q}^{(3)})^\top\|_{\mathrm{F}}=\|\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}},$$
$$\|\widetilde{\mathbf{N}}\|_{\mathrm{F}}=\|\mathbf{Q}^{(3)}\widetilde{\mathbf{\Lambda}}^{(3)}(\mathbf{Q}^{(3)})^\top\|_{\mathrm{F}}=\|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}.$$

We conclude that

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right)\leq\mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)+\|\mathbf{E}\|_{\mathrm{F}}\|\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}+\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}. \tag{49}$$

Next, we apply Lemma A.3 to $\mathbf{E}=\widehat{\mathbf{E}}$, $\mathbf{\Lambda}^{(3)}=\widehat{\mathbf{\Lambda}}^{(3)}$ and deduce that

$$\mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)\leq\sum_{l=1}^{d}\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\widetilde{\mathbf{\Lambda}}^{(3)}_{l,l}=\sum_{l=1}^{d}\Bigg(\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+\left(\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}-\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\right)\mathbf{\Lambda}^{(3)}_{l,l}$$
$$+\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\left(\widetilde{\mathbf{\Lambda}}^{(3)}_{l,l}-\mathbf{\Lambda}^{(3)}_{l,l}\right)\Bigg)$$
$$=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+\sum_{l=1}^{d}\left(\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}-\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\right)\mathbf{\Lambda}^{(3)}_{l,l}+\sum_{l=1}^{d}\widetilde{\mathbf{E}}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\left(\widetilde{\mathbf{\Lambda}}^{(3)}_{l,l}-\mathbf{\Lambda}^{(3)}_{l,l}\right)$$
$$=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+\mathrm{Trace}\left(\mathbf{\Pi}(\widetilde{\mathbf{E}}-\mathbf{E})\mathbf{\Pi}^\top\widetilde{\mathbf{\Lambda}}^{(3)}\right)+\mathrm{Trace}\left(\mathbf{\Pi}\widetilde{\mathbf{E}}\mathbf{\Pi}^\top(\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)})\right).$$

We apply Cauchy-Schwarz inequality again to the second and the third term:

$$\mathrm{Trace}\left(\mathbf{\Pi}(\widetilde{\mathbf{E}}-\mathbf{E})\mathbf{\Pi}^\top\widetilde{\mathbf{\Lambda}}^{(3)}\right)\leq\|(\widetilde{\mathbf{E}}-\mathbf{E})\mathbf{\Pi}^\top\|_{\mathrm{F}}\|\mathbf{\Pi}^\top\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}=\|\widetilde{\mathbf{E}}-\mathbf{E}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}},$$
$$\mathrm{Trace}\left(\mathbf{\Pi}\widetilde{\mathbf{E}}\mathbf{\Pi}^\top(\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)})\right)\leq\|\widetilde{\mathbf{E}}\mathbf{\Pi}^\top\|_{\mathrm{F}}\|\mathbf{\Pi}^\top(\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)})\|_{\mathrm{F}}=\|\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}$$

where we use invariance of the Frobenius norm under column and row permutations. We conclude that

$$\mathrm{Trace}\left(\widetilde{\mathbf{E}}\mathbf{Q}^{**}\widetilde{\mathbf{N}}(\mathbf{Q}^{**})^\top\right)\leq\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+\|\widetilde{\mathbf{E}}-\mathbf{E}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}+\|\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}.$$

We combine this inequality with (49) and obtain:

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right)\leq\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+2\|\mathbf{E}\|_{\mathrm{F}}\|\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}+2\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}. \tag{50}$$

Next, we use triangle inequality and deduce that

$$\|\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}\leq\|\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}+\|\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}.$$

Hence, we continue (50):

$$\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^\top\right)\leq\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+2\|\mathbf{E}\|_{\mathrm{F}}\|\mathbf{\Lambda}^{(3)}$$

$$-\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}+2\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\left(\|\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}+\|\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}\right)$$

$$=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+2\|\mathbf{E}\|_{\mathrm{F}}\|\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}+2\|\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}+2\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}$$

$$\leq\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+2\|\mathbf{E}\|_{\mathrm{F}}\|\mathbf{\Lambda}^{(3)}-\widetilde{\mathbf{\Lambda}}^{(3)}\|_{\mathrm{F}}+2\|\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}+2\|\mathbf{E}\|_{\mathrm{F}}\|\widetilde{\mathbf{\Lambda}}^{(3)}-\mathbf{\Lambda}^{(3)}\|_{\mathrm{F}}$$

where in the last transition we use $\|\mathbf{E}-\widetilde{\mathbf{E}}\|_{\mathrm{F}}\leq\|\mathbf{E}\|_{\mathrm{F}}$ which is according to (47). We continue this chain of inequalities using (47) again:

$$\begin{aligned}\mathrm{Trace}\left(\mathbf{E}\mathbf{Q}^{**}\mathbf{N}(\mathbf{Q}^{**})^{\top}\right)&\leq\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+\frac{2}{12}\delta+\frac{2}{12}\delta+\frac{2}{12}\delta=\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+\frac{\delta}{2}\\&<\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}_l,\boldsymbol{\pi}_l}\mathbf{\Lambda}^{(3)}_{l,l}+\delta.\end{aligned}$$

This is a contradiction with (46) taking into account $\mathbf{Q}^{**}$'s definition (48). Hence, (44) is proven. □

Let $\mathbf{Q}^{*}$ be defined as in Lemma A.4's statement. Further, we denote $\boldsymbol{\pi}(\mathbf{A})=\boldsymbol{\pi}$, $\mathbf{\Pi}(\mathbf{A})=\mathbf{\Pi}$ where $\boldsymbol{\pi},\mathbf{\Pi}$ are defined as in Lemma A.4's statement. That is, $\boldsymbol{\pi}(\mathbf{A})$ denotes some permutation which sorts diagonal entries of $\mathbf{E}$ in a non-ascending order. It's a function of $\mathbf{A}$ since $\mathbf{E}$ is a function of $\mathbf{A}$ defined in (35). In fact, based on (35), we see that $\boldsymbol{\pi}(\mathbf{A})$ is some permutation which sorts diagonal entries of $\mathbf{A}$ in a non-descending order. $\mathbf{\Pi}(\mathbf{A})$ denotes a permutation matrix corresponding to $\boldsymbol{\pi}(\mathbf{A})$. That is, diagonal entries of $\mathbf{\Pi}(\mathbf{A})\mathbf{A}\mathbf{\Pi}(\mathbf{A})^{\top}$ are sorted in a non-descending order.

Let $\mathcal{G}(\mathbf{A})$ denote the right hand side of (37) where we substitute $\mathbf{Q}=\mathbf{Q}^{*}$. That is, $\mathcal{G}(\mathbf{A})$ is an optimal value of $\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{SDE}};\mathcal{X},\mathcal{Y},\mathcal{T}_{\mathrm{SDE}})$ with $\mathbf{A}$ fixed:

$$\begin{aligned}\mathcal{G}(\mathbf{A})&=\log\det(\mathbf{I}_d-4\mathbf{A})-\frac{1}{2}\log\det(\mathbf{I}_d-8\mathbf{A})-L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2-L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2\\&\quad-2\sum_{l=1}^{d}\mathbf{E}_{\boldsymbol{\pi}(\mathbf{A})_l,\boldsymbol{\pi}(\mathbf{A})_l}\mathbf{\Lambda}^{(3)}_{l,l}\\&=\log\det(\mathbf{I}_d-4\mathbf{A})-\frac{1}{2}\log\det(\mathbf{I}_d-8\mathbf{A})-L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2-L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2\\&\quad+\sum_{l=1}^{d}\left(1+(1-8\mathbf{A}_{\boldsymbol{\pi}(\mathbf{A})_l,\boldsymbol{\pi}(\mathbf{A})_l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{l,l}\end{aligned}\tag{51}$$

where we use $\mathbf{E}$'s definition (35). Let $\boldsymbol{\pi}^{-1}(\mathbf{A})\in\mathbb{N}^d$ denote a permutation inverse to $\boldsymbol{\pi}(\mathbf{A})$. By rearranging terms in the sum, we have:

$$\sum_{l=1}^{d}\left(1+(1-8\mathbf{A}_{\boldsymbol{\pi}(\mathbf{A})_l,\boldsymbol{\pi}(\mathbf{A})_l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{l,l}=\sum_{l=1}^{d}\left(1+(1-8\mathbf{A}_{l,l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}.$$

Therefore, we have:

$$\begin{aligned}\mathcal{G}(\mathbf{A})&=\log\det(\mathbf{I}_d-4\mathbf{A})-\frac{1}{2}\log\det(\mathbf{I}_d-8\mathbf{A})-L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2-L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2\\&\quad+\sum_{l=1}^{d}\left(1+(1-8\mathbf{A}_{l,l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l,\boldsymbol{\pi}^{-1}(\mathbf{A})_l}.\end{aligned}\tag{52}$$

Define a new function $\mathcal{G}(\widehat{\mathbf{A}}, \mathbf{A})$, where $\widehat{\mathbf{A}} \in \mathbb{D}_d$ satisfies $8\widehat{\mathbf{A}} \prec \mathbf{I}_d$, as follows:

$$\mathcal{G}(\widehat{\mathbf{A}}, \mathbf{A}) = \log\det(\mathbf{I}_d - 4\widehat{\mathbf{A}}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\widehat{\mathbf{A}}) + \sum_{l=1}^{d}(1 - 8\widehat{\mathbf{A}}_{l,l})^{-1}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l}.$$

By the definition of $\mathcal{G}(\widehat{\mathbf{A}}, \mathbf{A})$, we have:

$$\mathcal{G}(\mathbf{A}) = \mathcal{G}(\mathbf{A}, \mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 + \sum_{l=1}^{d}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l}.$$

Hence, it holds:

$$\mathcal{G}(\mathbf{A}) \geq \inf_{\widehat{\mathbf{A}} \in \mathbb{D}_d,\, 8\widehat{\mathbf{A}} \prec \mathbf{I}_d} \mathcal{G}(\widehat{\mathbf{A}}, \mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 + \sum_{l=1}^{d}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l}. \quad (53)$$

Next, we show that there is a closed-form expression for the solution of $\inf_{\widehat{\mathbf{A}} \in \mathbb{D}_d,\, 8\widehat{\mathbf{A}} \prec \mathbf{I}_d} \mathcal{G}(\widehat{\mathbf{A}}, \mathbf{A})$. Since $\widehat{\mathbf{A}} \in \mathbb{D}_d$, we have: $\log\det(\mathbf{I}_d - 4\widehat{\mathbf{A}}) = \sum_{l=1}^{d}\log(1 - 4\widehat{\mathbf{A}}_{l,l})$, $\log\det(\mathbf{I}_d - 8\widehat{\mathbf{A}}) = \sum_{l=1}^{d}\log(1 - 8\widehat{\mathbf{A}}_{l,l})$. We further have:

$$\mathcal{G}(\widehat{\mathbf{A}}, \mathbf{A}) = \sum_{l=1}^{d}\left(\log(1 - 4\widehat{\mathbf{A}}_{l,l}) - \frac{1}{2}\log(1 - 8\widehat{\mathbf{A}}_{l,l}) + (1 - 8\widehat{\mathbf{A}}_{l,l})^{-1}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l}\right). \quad (54)$$

From (54), we see that minimization $\inf_{\widehat{\mathbf{A}} \in \mathbb{D}_d,\, 8\widehat{\mathbf{A}} \prec \mathbf{I}_d} \mathcal{G}(\widehat{\mathbf{A}}, \mathbf{A})$ reduces to d independent minimization problems with respect to $\widehat{\mathbf{A}}_{l,l}$ such that $8\widehat{\mathbf{A}}_{l,l} < 1$. l'th problem, $1 \leq l \leq d$, is solved using Lemma A.1 where we set $\phi = \mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l}$. Let $\mathbf{A}^{**} \in \mathbb{D}_d$ denote the corresponding solution. Then, for all $1 \leq l \leq d$, we have:

$$\mathbf{A}^{**}_{l,l} = \frac{1}{16}\left(1 - 2\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l} - \sqrt{\left(2\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l} + 1\right)^2 + 8\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l}}\right). \quad (55)$$

From (53) it follows that

$$\begin{aligned}
\mathcal{G}(\mathbf{A}) &\geq \mathcal{G}(\mathbf{A}^{**}, \mathbf{A}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 + \sum_{l=1}^{d}\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l} \\
&= \log\det(\mathbf{I}_d - 4\mathbf{A}^{**}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}^{**}) - L^{-1}\sum_{i=1}^{L}\|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^{L}\|\mathbf{y}^{(j)}\|^2 \\
&+ \sum_{l=1}^{d}\left(1 + (1 - 8\mathbf{A}^{**}_{l,l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l}. \quad (56)
\end{aligned}$$

Denote $\mathbf{E}^{**} = -\frac{1}{2}\mathbf{I}_d - \frac{1}{2}(\mathbf{I}_d - 8\mathbf{A}^{**})^{-1}$. Then we have:

$$\begin{aligned}
\sum_{l=1}^{d}\left(1 + (1 - 8\mathbf{A}^{**}_{l,l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{\boldsymbol{\pi}^{-1}(\mathbf{A})_l, \boldsymbol{\pi}^{-1}(\mathbf{A})_l} &= -2\mathrm{Trace}\left(\mathbf{E}^{**}\mathbf{\Pi}(\mathbf{A})^{-1}\mathbf{\Lambda}^{(3)}\left(\mathbf{\Pi}(\mathbf{A})^{-1}\right)^{\top}\right) \\
&\leq -2\sum_{l=1}^{d}\mathbf{E}^{**}_{\boldsymbol{\pi}(\mathbf{A}^{**})_l, \boldsymbol{\pi}(\mathbf{A}^{**})_l}\mathbf{\Lambda}^{(3)}_{l,l} \quad (57)
\end{aligned}$$

where the second transition follows from Lemma A.4 and the fact that $\boldsymbol{\pi}(\mathbf{A}^{**})$ sorts diagonal entries of $\mathbf{A}^{**}$ in a non-descending order, hence its sorts diagonal entries of $\mathbf{E}^{**}$ in a non-ascending order (recall the definition of $\boldsymbol{\pi}(\mathbf{A}^{**})$ and $\mathbf{E}^{**}$). Denote $\mathbf{E}^{*} = \mathbf{\Pi}(\mathbf{A}^{**})\mathbf{E}^{**}\mathbf{\Pi}(\mathbf{A}^{**})^{\top}$. Then $\mathbf{E}^{**}_{\boldsymbol{\pi}(\mathbf{A}^{**})_l, \boldsymbol{\pi}(\mathbf{A}^{**})_l} = \mathbf{E}^{*}_{l,l}$ for all $1 \leq l \leq d$ and

$$\sum_{l=1}^{d}\mathbf{E}^{**}_{\boldsymbol{\pi}(\mathbf{A}^{**})_l, \boldsymbol{\pi}(\mathbf{A}^{**})_l}\mathbf{\Lambda}^{(3)}_{l,l} = \sum_{l=1}^{d}\mathbf{E}^{*}_{l,l}\mathbf{\Lambda}^{(3)}_{l,l}. \quad (58)$$

Further, we have:

$$\begin{aligned}\mathbf{E}^* &= \mathbf{\Pi}(\mathbf{A}^{**})\left(-\frac{1}{2}\mathbf{I}_d - \frac{1}{2}(\mathbf{I}_d - 8\mathbf{A}^{**})^{-1}\right)\mathbf{\Pi}(\mathbf{A}^{**})^\top \\ &= -\frac{1}{2}\mathbf{I}_d - \frac{1}{2}(\mathbf{I}_d - 8\mathbf{\Pi}(\mathbf{A}^{**})\mathbf{A}^{**}\mathbf{\Pi}(\mathbf{A}^{**})^\top)^{-1} \\ &= -\frac{1}{2}\mathbf{I}_d - \frac{1}{2}(\mathbf{I}_d - 8\mathbf{A}^*)^{-1}\end{aligned}$$

where we denote $\mathbf{A}^* = \mathbf{\Pi}(\mathbf{A}^{**})\mathbf{A}^{**}\mathbf{\Pi}(\mathbf{A}^{**})^\top$, i.e. $\mathbf{A}^*_{l,l} = \mathbf{A}^{**}_{\boldsymbol{\pi}(\mathbf{A}^{**})_l,\boldsymbol{\pi}(\mathbf{A}^{**})_l}$ for all $1 \le l \le d$. Given the definition of $\mathbf{A}^{**}$ (55), for all $1 \le l \le d$ we have:

$$\mathbf{A}^*_{l,l} = \frac{1}{16}\left(1 - 2\mathbf{\Lambda}^{(3)}_{l,l} - \sqrt{\left(2\mathbf{\Lambda}^{(3)}_{l,l} + 1\right)^2 + 8\mathbf{\Lambda}^{(3)}_{l,l}}\right). \tag{59}$$

That is, $\mathbf{A}^*$ is independent of $\mathbf{A}$. Based on (59), we see that smaller values of $\mathbf{\Lambda}^{(3)}_{l,l}$ result in bigger values of $\mathbf{A}^*_{l,l}$. Since $\mathbf{\Lambda}^{(3)}_{1,1}, \ldots, \mathbf{\Lambda}^{(3)}_{d,d}$ are ordered in a non-ascending order, we deduce that $\mathbf{A}^*_{1,1}, \ldots, \mathbf{A}^*_{d,d}$ are ordered in a non-descending order. By the definition of $\boldsymbol{\pi}(\mathbf{A}^*)$, we then have $\mathbf{A}^*_{l,l} = \mathbf{A}^*_{\boldsymbol{\pi}(\mathbf{A}^*)_l,\boldsymbol{\pi}(\mathbf{A}^*)_l}$ for all $1 \le l \le d$. Therefore,

$$\sum_{l=1}^d \mathbf{E}^*_{l,l}\mathbf{\Lambda}^{(3)}_{l,l} = -\frac{1}{2}\sum_{l=1}^d (1 + (1 - 8\mathbf{A}^*_{l,l})^{-1})\mathbf{\Lambda}^{(3)}_{l,l} = -\frac{1}{2}\sum_{l=1}^d (1 + (1 - 8\mathbf{A}^*_{\boldsymbol{\pi}(\mathbf{A}^*)_l,\boldsymbol{\pi}(\mathbf{A}^*)_l})^{-1})\mathbf{\Lambda}^{(3)}_{l,l}.$$

Combining this with (58), (57), we can continue the chain of inequalities (56):

$$\begin{aligned}\mathcal{G}(\mathbf{A}) &\ge \log\det(\mathbf{I}_d - 4\mathbf{A}^{**}) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}^{**}) - L^{-1}\sum_{i=1}^L \|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^L \|\mathbf{y}^{(j)}\|^2 \\ &\quad + \sum_{l=1}^d \left(1 + (1 - 8\mathbf{A}^*_{\boldsymbol{\pi}(\mathbf{A}^*)_l,\boldsymbol{\pi}(\mathbf{A}^*)_l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{l,l} \\ &= \log\det(\mathbf{I}_d - 4\mathbf{A}^*) - \frac{1}{2}\log\det(\mathbf{I}_d - 8\mathbf{A}*) - L^{-1}\sum_{i=1}^L \|\mathbf{x}^{(i)}\|^2 - L^{-1}\sum_{j=1}^L \|\mathbf{y}^{(j)}\|^2 \\ &\quad + \sum_{l=1}^d \left(1 + (1 - 8\mathbf{A}^*_{\boldsymbol{\pi}(\mathbf{A}^*)_l,\boldsymbol{\pi}(\mathbf{A}^*)_l})^{-1}\right)\mathbf{\Lambda}^{(3)}_{l,l} = \mathcal{G}(\mathbf{A}^*)\end{aligned} \tag{60}$$

where in the second transition we use the fact that

$$\det(\mathbf{I}_d - 4\mathbf{A}^*) = \det\left(\mathbf{\Pi}(\mathbf{A}^{**})(\mathbf{I}_d - 4\mathbf{A}^{**})\mathbf{\Pi}(\mathbf{A}^{**})^\top\right) = \det\left(\mathbf{I}_d - 4\mathbf{A}^{**}\right)$$

and, similarly, $\det(\mathbf{I}_d - 8\mathbf{A}^*) = \det(\mathbf{I}_d - 8\mathbf{A}^{**})$. In the third transition, we use definition of $\mathcal{G}(\cdot)$ (51). Note that (60) holds for all $\mathbf{A} \in \mathbb{D}_d$ such that $8\mathbf{A} \prec \mathbf{I}_d$ and also $8\mathbf{A}^* \prec \mathbf{I}_d$ since $8\mathbf{A}^{**} \prec \mathbf{I}_d$. We conclude that, when $\mathbf{B}, \mathbf{C}, D$ are chosen optimally with a given $\mathbf{A}$, the minimum of $\overline{\mathcal{L}}(\boldsymbol{\theta}_{\mathrm{SDE}}; \mathcal{X}, \mathcal{Y}, \mathcal{T}_{\mathrm{SDE}})$ is reached when $\mathbf{A} = \mathbf{A}^*$. As we have already deduced, diagonal entries of $\mathbf{A} = \mathbf{A}^*$ are sorted in the non-descending order. Hence, using Lemma A.4's notation, diagonal entries of $\mathbf{E}$ are already sorted in a non-ascending sorting order and $\boldsymbol{\pi} = (1, \ldots, d)$, $\mathbf{\Pi} = \mathbf{I}_d$ satisfy requirements of the Lemma. Hence, with $\mathbf{A} = \mathbf{A}^*$, the optimal $\mathbf{B}$ has a form $(\mathbf{I}_d - 4\mathbf{A})^{1/2}\mathbf{Q}^*$ where $\mathbf{Q}^* = \mathbf{I}_d(\mathbf{Q}^{(3)})^\top = (\mathbf{Q}^{(3)})^\top$. Optimal $\mathbf{C}$ and D are further determined by (33). (11) follows from (60) and the fact that, as discussed above, we can replace $\boldsymbol{\pi}(\mathbf{A}^*)_l$ with l in (60), $1 \le l \le d$. □

B Additional experimental details

B.1 Compute resources and implementation

We use NumPy [24] in Google Colaboratory the variance comparison and kernel classification experiment. For the Transformer setups, we use TPU cluster and JAX [7] library. All tested Transformer variants were trained and tested on a TPU pods containing 4 TPU v3 chips with JAX and on GPUs (V100).

B.2 Variance comparison

We repeat the setup of [30] closely: we draw 5 pairs of sets $\{\mathbf{x}^{(i)}\}_{1\leq i\leq L}$, $\{\mathbf{y}^{(j)}\}_{1\leq j\leq L}$, $L = 1024$. On each pair, we compute the relative variance for all pairs of points and for all indicated RF methods. Further, the shifted log-variance is optimized separately on each pair of sets for GERF, ADERF and SDERF.

We take $M = 1$ since M's value is not important in this experiment: bigger M would just shift the curves below. The reported curves are means over all pairs of points and over all 5 sets.

B.3 Kernel classification

As in [30], we obtain training, validation and test splits by shuffling the raw dataset and taking 90%, 5%, 5% objects respectively. The splits are fixed for all RF methods. We tune σ on a logarithmic grid of 10 values on $[10^{-2}, 10^{2}]$. For each σ and each RF type, we try 50 seeds for drawing RFs during validation and testing. Testing is performed for the best σ only. Figure 3 reports averages over 50 seeds. We use orthogonal ω's for all types of RFs as described in [30], since orthogonal random features work better in practice [15, 30].

B.4 DERFs for long-sequence Transformers

B.4.1 Speech modelling

Our Conformer-Transducer variant was characterized by: **20** conformer layers, $\mathrm{model\_dim} = \mathbf{512}$, relative position embedding dimensionality $\mathrm{rped} = \mathbf{512}$ and $h = \mathbf{8}$ heads. We used batch size $\mathrm{bs} = \mathbf{2048}$ and trained with the adam optimizer on TPUs. For the regular Conformer-Transducer training, we run ablation studies over different number of random features: $m = \mathbf{8}, \mathbf{32}, \mathbf{128}$. In the NST setting, we run experiments with $m = \mathbf{8}$. We reported commonly used metric: normalized word error rate (NWER).

B.4.2 Natural language processing

We pretrained BERT model on two publicly available datasets (see: Table 3). Following the original BERT training, we mask 15% of tokens in these two datasets, and train to predict the mask. All methods were warm started from exactly the same pre-trained checkpoint after 1M iteration of BERT pretraining. We used the exact same hyperparameter-setup for all the baselines (FAVOR++[30], FAVOR+ [15], ELU [25], ReLU [15]) and FAVOR++. The hyperparameters for pretraining are shown in Table 2. We finetuned on GLUE task, warm-starting with the weights of the pretrained model. The setup is analogous to the one from the original BERT paper.

Table 2: Hyperparameters for the base models for pre-training for all methods

Parameter	Value
# of heads	12
# of hidden layers	12
Hidden layer size	768
# of tokens	512
Batch size	256
M	256
Pretrain Steps	$1M$
Loss	MLM
Activation layer	gelu
Dropout prob	0.1
Attention dropout prob	0.1
Optimizer	Adam
Learning rate	10^{-4}
Compute resources	8×8 TPUv3

Table 3: Dataset used for pre training.

Dataset	# tokens	Avg. doc len.
Books [52]	1.0B	37K
Wikipedia	3.1B	592