A Proofs from Section 4

We collect together proofs and auxiliary algorithms from Section 4.

A.1 Proof of Lemma 4.1

Lemma A.1. *The exponential of a matrix in the form* $\begin{pmatrix} a\mathbf{I}_n & \mathbf{u} \\ \mathbf{u}^\top & b \end{pmatrix}$, *where a and b are nonnegative, is*

$$e^{\phi}\begin{pmatrix} (\cosh\psi+\sinh\psi\cos\gamma)\hat{\mathbf{u}}\hat{\mathbf{u}}^\top & \sinh\psi\sin\gamma\hat{\mathbf{u}} \\ \sinh\psi\sin\gamma\hat{\mathbf{u}}^\top & \cosh\psi-\sinh\psi\cos\gamma \end{pmatrix}+e^{a}\begin{pmatrix} \mathbf{I}_n-\hat{\mathbf{u}}\hat{\mathbf{u}}^\top & 0 \\ 0 & 0 \end{pmatrix}, \tag{A.1}$$

where $\hat{\mathbf{u}}$ *is the unit vector* $\mathbf{u}/\|\mathbf{u}\|$, $\phi=(a+b)/2$, $\psi=\sqrt{(a-b)^2/4+\|\mathbf{u}\|^2}$, *and* $\gamma=\tan^{-1}(2\|\mathbf{u}\|/(a-b))$.

We symbolically exponentiate an $n+1\times n+1$ matrix of the form

$$\mathbf{M}=\begin{pmatrix} aI_n & \mathbf{u} \\ \mathbf{u}^\top & b \end{pmatrix}.$$

Since this matrix is real and symmetric, its eigenvalues λ_i are positive and its unit eigenvectors $\mathbf{v}_i$ form an orthonormal basis. The method that we use to symbolically exponentiate it is to express it in the form

$$\mathbf{M}=\sum_{i=1}^{n}\lambda_i\mathbf{v}_i\mathbf{v}_i^\top.$$

The exponential then becomes

$$e^{\mathbf{M}}=\sum_{i=1}^{n}e^{\lambda_i}\mathbf{v}_i\mathbf{v}_i^\top.$$

As a matter of notation, let $\hat{\mathbf{u}}$ be the unit vector such that $\|\mathbf{u}\|\hat{\mathbf{u}}=\mathbf{u}$.

Eigenvalues. The characteristic equation for $\mathbf{M}$ is not difficult to calculate. It is:

$$(\lambda-a)^{n-1}(\lambda^2-(a+b)\lambda+ab-\|\mathbf{u}\|^2). \tag{A.2}$$

This yields $n-1$ eigenvalues equal to a, and the other two equal to $(a+b)/2+\sqrt{(a-b)^2/4+\|\mathbf{u}\|^2}$ and $(a+b)/2-\sqrt{(a-b)^2/4+\|\mathbf{u}\|^2}$. We label them λ_1 and λ_2, respectively, and the rest are equal to a.

Eigenvectors. First we show that $\mathbf{M}$ has two eigenvectors of the form $(\mathbf{u},c)^\top$:

$$\begin{pmatrix} aI_n & \mathbf{u} \\ \mathbf{u}^\top & b \end{pmatrix}\begin{pmatrix} \mathbf{u} \\ c \end{pmatrix}=\begin{pmatrix} (a+c)\mathbf{u} \\ \|\mathbf{u}\|^2+bc \end{pmatrix},$$

So as long as we choose c such that $c^2+ac=\|\mathbf{u}\|^2+bc$, or $c=(b-a)/2\pm\sqrt{(a-b)^2/4+\|\mathbf{u}\|^2}$, then $(\mathbf{u},c)^\top$ is an eigenvector with eigenvalue $a+c$. These two eigenvalues are just λ_1 and λ_2. We will call the corresponding eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$.

Since $\mathbf{M}$ is symmetric, all of its eigenvectors are orthogonal. The remaining eigenvectors are of the form $(\mathbf{w},0)^\top$, where $\mathbf{w}^\top\mathbf{u}=0$:

$$\begin{pmatrix} aI_n & \mathbf{u} \\ \mathbf{u}^\top & b \end{pmatrix}\begin{pmatrix} \mathbf{w} \\ 0 \end{pmatrix}=\begin{pmatrix} a\mathbf{w} \\ 0 \end{pmatrix}.$$

Clearly the corresponding eigenvalue for any such eigenvector is a, so there are $n-1$ of them. The corresponding parts of these eigenvectors are labeled $\mathbf{w}_i$, where $3\le i\le n+1$, and we assume they are unit vectors.

Since

$$e^{\mathbf{M}} = \sum_{i=1}^{n} e^{\lambda_i} \mathbf{v}_i \mathbf{v}_i^\top,$$

and the eigenvalue a is of multiplicity $n-1$, we have

$$\begin{aligned} e^{\mathbf{M}} &= e^{\lambda_1} \frac{\mathbf{v}_1 \mathbf{v}_1^\top}{\|\mathbf{v}_1\|^2} + e^{\lambda_2} \frac{\mathbf{v}_2 \mathbf{v}_2^\top}{\|\mathbf{v}_2\|^2} + e^a \sum_{i=3}^{n} \begin{pmatrix} \mathbf{w}_i \mathbf{w}_i^\top & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{pmatrix} \\ &= \frac{e^{\lambda_1}}{\|\mathbf{u}\|^2 + c_1^2} \begin{pmatrix} \mathbf{u}\mathbf{u}^\top & c_1 \mathbf{u} \\ c_1 \mathbf{u}^\top & c_1^2 \end{pmatrix} + \frac{e^{\lambda_2}}{\|\mathbf{u}\|^2 + c_2^2} \begin{pmatrix} \mathbf{u}\mathbf{u}^\top & c_2 \mathbf{u} \\ c_2 \mathbf{u}^\top & c_2^2 \end{pmatrix} + e^a \begin{pmatrix} \mathbf{I}_n - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{pmatrix} \end{aligned}$$

The last term in the equality is due to the fact that $\hat{\mathbf{u}}$ and the $\hat{\mathbf{w}}_i$ form an orthonormal basis for $\mathbb{R}^n$, so $\hat{\mathbf{u}}\hat{\mathbf{u}}^\top + \sum \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^\top = I_n$. We can reduce some of the factors in the expression by observing that $\|\mathbf{u}\| = -c_1 c_2$. Let

$$\begin{aligned} \beta_1^2 &= \|\mathbf{u}\|^2/(\|\mathbf{u}\|^2 + c_1^2) = -c_1 c_2/(c_1^2 - c_1 c_2) = c_2/(c_2 - c_1) = c_2^2/(c_2^2 - c_1 c_2) = c_2^2/(\|\mathbf{u}\|^2 + c_2^2) \\ \beta_2^2 &= \|\mathbf{u}\|^2/(\|\mathbf{u}\|^2 + c_2^2) = -c_1 c_2/(c_2^2 - c_1 c_2) = -c_1/(c_2 - c_1) = c_1^2/(c_1^2 - c_1 c_2) = c_1^2/(\|\mathbf{u}\|^2 + c_1^2) \end{aligned}$$

Note also that $\beta_1^2 + \beta_2^2 = 1$.

The Exponential. All that remains is to put everything together:

$$\begin{aligned} e^M &= \sum_{i=1}^{n} e^{\lambda_i} \mathbf{v}_i \mathbf{v}_i^\top \\ &= e^{\lambda_1} \begin{pmatrix} \beta_1^2 \hat{\mathbf{u}}\hat{\mathbf{u}}^\top & \beta_1 \beta_2 \hat{\mathbf{u}} \\ \beta_1 \beta_2 \hat{\mathbf{u}}^\top & \beta_2^2 \end{pmatrix} + e^{\lambda_2} \begin{pmatrix} \beta_2^2 \hat{\mathbf{u}}\hat{\mathbf{u}}^\top & -\beta_1 \beta_2 \hat{\mathbf{u}} \\ -\beta_1 \beta_2 \hat{\mathbf{u}}^\top & \beta_1^2 \end{pmatrix} + e^a \begin{pmatrix} I_n - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{pmatrix}. \end{aligned}$$

Some variable substitutions will give us the form in (A.1); $\lambda_1 = \phi + \psi$, $\lambda_2 = \phi - \psi$, and $\beta_1 = \cos(\gamma/2)$:

$$= e^{\phi} \begin{pmatrix} (\cosh\psi + \sinh\psi\cos\gamma)\hat{\mathbf{u}}\hat{\mathbf{u}}^\top & \sinh\psi\sin\gamma\hat{\mathbf{u}} \\ \sinh\psi\sin\gamma\hat{\mathbf{u}}^\top & \cosh\psi - \sinh\psi\cos\gamma \end{pmatrix} + e^a \begin{pmatrix} I_n - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top & 0 \\ 0 & 0 \end{pmatrix}.$$

A.2 Selecting α

Recall that $\mathbf{u}_i = \mathbf{A}_i \alpha$ (from section 4). Let $\hat{\mathbf{u}}_i = \mathbf{u}_i/\|\mathbf{u_i}\|$. Let us denote the elements of the matrix in (A.1) as

$$\begin{aligned} p_i^{11} &= e^{\phi}(\cosh\psi + \sinh\psi\cos\gamma) && = e^{\phi}\cosh\psi \\ p_i^{12} &= e^{\phi}(\sinh\psi\sin\gamma) && = -e^{\phi}\sinh\psi \\ p_i^{22} &= e^{\phi}(\cosh\psi - \sinh\psi\cos\gamma) && = e^{\phi}\cosh\psi \end{aligned}$$

We observe that $\hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top$ is a rank one outer product with unit trace and that a_i is the same for all i. So $\sum_i Q_i(\alpha) \bullet \mathbf{P}_i$ is given as:

$$\left(\sum_{i=0}^{m} (2 p_i^{12} \hat{\mathbf{u}}_i^\top \mathbf{A}_i) \right) \alpha \geq -m(n-1)e^a - \sum_{i=0}^{m} (p_i^{11} + p_i^{22} s). \tag{A.3}$$

It is worth noting that the right hand side bears a close resemblance to the trace of $\mathbf{P}$, which is $m(n-1)e^a + \sum_{i=0}^{m}(p_i^{11} + p_i^{22})$ (the trace of $\mathbf{I}_n$ is n, and the trace of $\hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top$ is always 1). We know that $s = \omega = 1$ (see Section 4), so this makes the RHS equal to the trace. Also from normalization step of Algorithm 1 we know that $\mathbf{P}$ is normalized with trace 1, so we have the following:

$$\left(\sum_{i=1}^{m} (2 p_i^{12} \hat{\mathbf{u}}_i^\top \mathbf{A}_i) \right) \alpha^{(t)} \geq -1. \tag{A.4}$$

Practical Considerations. We highlight two important practical consequences of our formulation. First, the procedure produces a very sparse update to α: in each iteration, only two coordinates of α are updated. This makes each iteration very efficient, taking only linear time. Second, by expressing $\mathbf{u}_i$ in terms of $\mathbf{g}_i$ we never need to explicitly compute $\mathbf{A}_i$ (as $\mathbf{u}_i = \mathbf{A}_i\alpha$), which in turn means that we do not need to compute the (expensive) square root of $\mathbf{G_i}$ explicitly.

Another beneficial feature of the dual-finding procedure for MKL is that terms involving the primal variables $\mathbf{P}$ are either normalized (when we set the trace of $\mathbf{P}$ to 1) or eliminated (due to the fact that we have a compact closed-form expression for $\mathbf{P}$), *which means that we never have to explicitly maintain* $\mathbf{P}$, save for a small number ($4m$) of variables.

A.3 Proof that ρ^2 is $O(1)$

Lemma A.2. ρ *is bounded by* $3/2$.

Proof. ρ is defined as the maximum of $\|\mathbf{Q}(\alpha^{(t)})\|$ for all t. Here $\|\cdot\|$ denotes the largest eigenvalue in absolute value [3]. Because $s = \omega = 1$ (see Section 4), the eigenvalues of $\mathbf{Q}_i(\alpha^{(t)})$ are 1 (with multiplicity $n-1$), and $1 \pm \|\mathbf{A}_i\alpha^{(t)}\|$. The greater of these in absolute value is clearly $1 + \|\mathbf{A}_i\alpha^{(t)}\|$.

$\|\mathbf{A}_i\alpha^{(t)}\|$ is equal to

$$((\alpha^{(t)})^T\mathbf{A}_i^T\mathbf{A}_i\alpha^{(t)})^{\frac{1}{2}} = \left(\frac{1}{r_i}(\alpha^{(t)})^T\mathbf{G}_i\alpha^{(t)}\right)^{\frac{1}{2}}.$$

$\alpha^{(t)}$ always has two nonzero elements, and they are equal to $1/2$. They also correspond to values of $\mathbf{y}$ with opposite signs, so if j and k are the coordinates in question, $(\alpha^{(t)})^T\mathbf{G}_i\alpha^{(t)} \leq (1/4)(\mathbf{G}_{i(jj)} + \mathbf{G}_{i(kk)})$, because $\mathbf{G}_{i(jk)}$ and $\mathbf{G}_{i(kj)}$ are both negative. Because of the factor of $1/r_i$, and because r_i is the trace of $\mathbf{G}_i$, $\|\mathbf{A}_i\alpha^{(t)}\| \leq 1/2$. This is true for any of the i, so the maximum eigenvalue of $\mathbf{Q}(\alpha^{(t)})$ in absolute value is bounded by $1 + 1/2 = 3/2$. $\square$

A.4 Exponentiating M

From $\mathbf{M}_i^{(t)}$ in Algorithm 1 and (4.2), we have $\mathbf{M}_i^{(t)} = \frac{1}{2\rho}(\mathbf{Q}_i(\alpha^{(t)}) + \rho\mathbf{I}_{n+1})$, where ρ is a program parameter which is explained in 4.2.

Our $\mathbf{Q}_i(\alpha) = \begin{pmatrix} \mathbf{I}_n & \mathbf{A}_i\alpha \\ (\mathbf{A}_i\alpha)^\top & 1 \end{pmatrix}$ is of the form $\begin{pmatrix} a\mathbf{I}_n & \mathbf{u_i} \\ \mathbf{u_i}^\top & b \end{pmatrix}$, where $a = 1$ and $b = 1$ are non-negative $\forall i$ and $\mathbf{u}_i = \mathbf{A}_i\alpha$. So we have

$$\mathbf{u}_i^\top\mathbf{u}_i = (\mathbf{A}_i\alpha)^\top\mathbf{A}_i\alpha = \alpha^\top\mathbf{A}_i^\top\mathbf{A}_i\alpha = \alpha^\top\frac{1}{r_i}\mathbf{G}_i\alpha \tag{A.5}$$

where the last equality follows from $\mathbf{A}_i^\top\mathbf{A}_i = \frac{1}{r_i}\mathbf{G}_i$ (cf. (4.2)). As we shall show in Algorithm 3, at each iteration the matrix to be exponentiated is a sum of matrices of the form $\frac{1}{2\rho}(\mathbf{Q}_i(\sum_{t=1}^{\tau}\alpha^{(t)}) + \rho t\mathbf{I}_{n+1})$, so Lemma A.1 can be applied at every iteration. Additionally, $a = b$, so many of the substitutions simplify considerably: $\phi = a$, $\psi = \|\mathbf{u}\|$, $\sin\gamma = \pm 1$, and $\cos\gamma = 0$.

We provide in detail the algorithm we use to exponentiate the matrix $\mathbf{M}$. This subroutine is called from Algorithm 3 in Section 4.

Practical considerations. In Lemma A.1, large inputs to the functions exp, cosh, and sinh will cause them to rapidly overflow even at double-precision range. Fortunately there are two steps we can take. First, $\exp(x)/2$ gets exponentially close to both $\sinh(x)$ and $\cosh(x)$ as x gets larger, so above a high enough value, we can simply approximate $\sinh(x)$ and $\cosh(x)$ with $\exp(x)/2$.

Because exp can overflow just as much as sinh or cosh, this doesn't solve the problem completely. However, since $\mathbf{P}$ is always normalized so that $\mathrm{tr}(\mathbf{P}) = 1$, we can multiply the elements of $\mathbf{P}$ by any factor we choose and the factor will be normalized out in the end. So above a certain value, we can use exp alone and throw a "quashing" factor ($e^{-\phi-q}$) into the equations before computing the result, and it will be normalized out later in the computation. For our purposes, setting $q = 20$ suffices. Note that this trades overflow for underflow, but underflow can be interpreted merely as one kernel disappearing from significance.

Algorithm 4 EXPONENTIATE-M

Input: $\mathbf{y}$, α, $\{\mathbf{G}_i\}$, ε', ρ, t

  $\phi \leftarrow -\frac{\varepsilon'}{2\rho}(1+\rho)t$
  for $i \in [1..m]$ **do**
    $\|\mathbf{u}_i\| \leftarrow \sqrt{\alpha^T \mathbf{G}_i \alpha}$
    $\mathbf{g}_i \leftarrow \frac{1}{\|\mathbf{u}_i\|}\mathbf{G}_i\alpha$
    $\psi_i \leftarrow \frac{\varepsilon'}{2\rho}\|\mathbf{u}_i\|$
  end for
  $q \leftarrow \max_i \psi_i$
  if $q < 20$ **then**
    for $i \in [1..m]$ **do**
      $l_i^{11} \leftarrow \cosh(\psi_i)$
      $l_i^{12} \leftarrow -\sinh(\psi_i)$
    end for
    $e_{\mathbf{M}} \leftarrow 1$
  else
    for $i \in [1..m]$ **do**
      $l_i^{11} \leftarrow e^{\psi_i - q}$
      $l_i^{12} \leftarrow -l_i^{11}$
    end for
    $e_{\mathbf{M}} \leftarrow 2e^{-q}$
  end if
  $S \leftarrow m(n-1)e_{\mathbf{M}} + 2\sum_i l_i^{11}$
  for $i \in [1..m]$ **do**
    $l_i^{11} \leftarrow l_i^{11}/S$
    $l_i^{12} \leftarrow l_i^{12}/S$
  end for
  $\mathbf{g} \leftarrow \sum_i 2 l_i^{12} \mathbf{g}_i$
  Return $\mathbf{l_i^{12}}$, $\mathbf{g}$
