If Only My Posterior Were Normal: Introducing Fisher HMC

Adrian Seyboldt
adrian.seyboldt@gmail.com
PyMC Labs

2025-04-07

Abstract

Hamiltonian Monte Carlo (HMC) is a powerful tool for Bayesian inference, as it can explore complex and high-dimensional parameter spaces. But HMC's performance is highly sensitive to the geometry of the posterior distribution, which is often poorly approximated by traditional mass matrix adaptations, especially in cases of non-normal or correlated posteriors. We propose Fisher HMC, an adaptive framework that uses the Fisher divergence to guide transformations of the parameter space. It generalizes mass matrix adaptation from affine functions to arbitrary diffeomorphisms. By aligning the score function of the transformed posterior with those of a standard normal distribution, our method identifies transformations that adapt to the posterior's scale and shape. We develop theoretical foundations efficient implementation strategies, and demonstrate significant sampling improvements. Our implementation, nutpie, integrates with PyMC and Stan and delivers better efficiency compared to existing samplers.

Keywords Bayesian Inference · Hamiltonian Monte Carlo · Mass Matrix Adaptation · Normalizing Flows · Fisher Divergence

1 Introduction

Hamiltonian Monte Carlo (HMC) is a powerful Markov Chain Monte Carlo (MCMC) method widely used in Bayesian inference for sampling from complex posterior distributions. HMC can explore high-dimensional parameter spaces more efficiently than traditional MCMC techniques, which makes it popular in probabilistic programming libraries like Stan Carpenter et al. (2017) and PyMC. However, the performance of HMC depends critically on the parameterization of the posterior space. Modern samplers automate a part of these reparametrizations by adapting a "mass matrix" in the warmup phase of sampling. A common approach in HMC is to estimate a mass matrix based on the inverse of the posterior covariance, typically in a diagonalized form, to adjust for

differences in scale across dimensions. We can think of this as a reparametrization that simply rescales the parameters such that they have a posterior variance of one. It is not obvious, however, that this is the best rescaling that can be done. Morever, even a well-tuned mass matrix can not do much to help us when sampling from more challenging posterior distributions such as those with strong correlations in high dimensions, or with funnel-like pathologies. For researchers working with multilevel hierarchical models with correlated group-level parameters, manually rescaling and rotating the parameter space to improve sampling efficiency requires deep statistical expertise and can be time-consuming. In many cases, good reparametrizations are also data-dependent, which makes it difficult to write a model once, and apply it to a wide range of individual datasets.

To address these limitations, we propose an adaptive HMC framework that extends beyond the traditional concept of a mass matrix: instead of just rescaling variables, we allow for arbitrary diffeomorphisms that dynamically transform the parameter space. We use the Fisher divergence as a criterion to choose between different transformations, which allows us to adapt the geometry of the posterior space in a way that optimizes HMC's efficiency. By aligning the scores (derivatives of the log-density) of the transformed posterior with those of a standard normal distribution, we approximate an idealized parameterization that facilitates efficient sampling. In the first section, we motivate why the scores are useful for for Hamiltonian dynamics. We then present the Fisher divergence as a metric by which we can assess the transformations of target distributions, deriving closed-form solutions for optimal diffeomorphisms in the affine case, i.e. mass matrices. Finally, we suggest additional modifications to the adaptation schedule shared by the major software implementations, which complement gradient-based adaptation.

2 Fisher HMC: Motivation and Theory

2.1 Motivation: Example with Normal Posterior

HMC is a gradient-based method, meaning that the algorithm computes the derivatives of the log posterior density (the scores). While these gradients contain significant information about the target density, traditional methods of mass matrix adaptation ignore them. To illustrate how useful the scores can be, consider a standard normal posterior $N(\mu, \sigma^2)$ with density $p(x) \propto \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)$. Let's assume we have two posterior draws x_1 and x_2, together with the covector of scores

$$\alpha_i = \frac{\partial}{\partial x_i} \log p(x_i) = \sigma^{-2}(\mu - x_i). \tag{1}$$

Based on this information alone, we can directly compute μ and σ to identify the exact posterior. Solving for μ and σ, we get

$$\mu = \bar{x} + \sigma^2 \bar{\alpha} \quad \text{and} \quad \sigma^2 = \mathrm{Var}[x_i]^{\frac{1}{2}} \mathrm{Var}[\alpha_i]^{-\frac{1}{2}}, \tag{2}$$

where $\bar{x}$ and $\bar{\alpha}$ are the sample means of x_i and α_i, respectively. If we take advantage of the scores, we can compute the exact posterior and thus an optimal mass matrix with no sample variance, based on just two draws!

This generalizes directly to multivariate normal posteriors $N(\mu, \Sigma)$, where we can leverage the elegant fact that the scores are normally distributed with covariance Σ^{-1}. Assume we have $N+1$ linearly independent draws $x_i \in \mathbb{R}^N$ with scores $\alpha_i = \Sigma^{-1}(x_i - \mu)$. The mean of these equations gives us $\mu = \bar{x} - \Sigma\bar{\alpha}$. It follows that $\Sigma^{-1}X = S$, where the i-th column of S is $\alpha_i - \bar{\alpha}$, and the i-th column of X is $x_i - \bar{x}$. Finally, we have

$$SS^T = \mathrm{Cov}[\alpha_i] = \Sigma^{-1}XX^T\Sigma^{-1} = \Sigma^{-1}\,\mathrm{Cov}[x_i]\Sigma^{-1} \tag{3}$$

and we can recover Σ as the geometric mean of the positive symmetric matrices $\mathrm{Cov}[x_i]$ and $\mathrm{Cov}[s_i]^{-1}$:

$$\Sigma = \mathrm{Cov}[x_i]^{-\frac{1}{2}}\Big(\mathrm{Cov}[x_i]^{\frac{1}{2}}\,\mathrm{Cov}[\alpha_i]\mathrm{Cov}[x_i]^{\frac{1}{2}}\Big)^{\frac{1}{2}}\mathrm{Cov}[x_i]^{-\frac{1}{2}} \tag{4}$$

In this way we can compute the parameters of the normal distribution exactly. Of course, most posterior distributions of interest are not multivariate normal, and if they were, we would not have to run MCMC in the first place. But it is common in Bayesian inference for the posterior to approximate a normal distribution reasonably well, which suggests that the scores contain useful information that is ignored in standard methods.

2.2 Transformed HMC

When we manually reparameterize a model to make HMC more efficient, we try to find a transformation (or diffeomorphism) of our posterior such that HMC performs better on it. Formally, if our posterior μ is defined on a space M, we try to find a diffeomorphism $f : N \to M$ such that the transformed posterior $f^*\mu$ is well-behaved with respect to some property. Note that we define the transformation as a function *from* the transformed space *to* the original posterior, in keeping consistent with the Normalizing Flow literature. Since the transformation is a bijection, we can choose any direction we want, as long as we stay consistent with our choice. $f^*\mu$ refers to the pullback of the posterior (which we can interpret as a volume form), i.e. we *pull it back* to the space N along the transformation f. If f is an affine transformation, this simplifies to mass matrix-based HMC, wherein choosing $f(x) = \Sigma^{\frac{1}{2}}x + \mu$ corresponds to the mass matrix Σ^{-1}, as described in more detail in (Neal 2012). In the present work, we restrict ourselves to the subset of affine diffeomorphisms, meaning that the transformations we derive are implemented in HMC in the same way as previous mass matrix adaptative HMC schemas, in the leapfrog integrator.

HMC efficiency is notoriously dependent on the parametrization, so it is to be expected that transformed HMC be much more efficient for some choices of f than for others. It is not, however, obvious what criterion should be used to evaluate a particular choice of f, in order to guide an automatic learning of the transformation. We need a loss function that maps the diffeomorphism to a measure of difficulty for HMC. This is hard to quantify in general, but we can observe that HMC efficiency largely depends on the trajectory, which in fact does not depend on the density directly, but rather

only on the scores. Therefore, a reasonable loss function might assess how well the transformed space's *scores* align with those of our desired transformed posterior. We choose the standard normal distribution as the ideal transformed posterior, since we know that HMC is efficient in this case, given the nice Gaussian properties such as constant curvature. This still leaves open the choice of a specific norm for comparing the scores of the standard normal with those of the transformed posterior. But since the standard normal distribution is defined in terms of an inner product, we already have a well-defined norm on the scores that allows us to evaluate their difference. This directly motivates the following definition of the Fisher divergence.

2.3 Fisher divergence

Let (N, g) be a Riemannian manifold with probability volume forms ω_1 and ω_2. We can define a scalar function z on N by $\omega_2 = z\omega_1$, or equivalently we also write this as $z = \frac{\omega_2}{\omega_1}$.

We define the Fisher divergence of ω_1 and ω_2 as

$$\mathcal{F}_g(\omega_1, \omega_2) = \int \|\nabla \log(z)\|_g^2 d\omega_1. \tag{5}$$

Note that $\mathcal{F}$ requires more structure on N than KL-divergence $\int \log(z) d\omega_1$, as the norm depends on the metric tensor. Given a second (non-Riemannian) manifold M with a probability volume form μ, and a diffeomorphism $f : N \to M$, we can define the divergence between μ and ω_1 by pulling back μ to N, i.e. $\mathcal{F}_g(f^*\mu, \omega_1)$.

We can also compute this Fisher divergence directly on M, by pushing the metric tensor to M:

$$\mathcal{F}_g(f^*\mu, \omega_1) = \mathcal{F}_{(f^{-1})^* g}\big(\mu, (f^{-1})^* \omega_1\big) \tag{6}$$

In this case, μ is our posterior, M is the space on which it is originally defined, and ω_1 is the standard normal distribution.

2.4 Affine choices for the diffeomorphism F

We focus on three families of affine diffeomorphisms F, for which derive specific results.

2.4.1 Diagonal mass matrix

If we choose $F_{\sigma,\mu} : Y \to X$ as $x \mapsto y \odot \sigma + \mu$, we are effectively doing diagonal mass matrix estimation. In this case, the sample Fisher divergence reduces to

$$\hat{F}_{\sigma,\mu}(f^*Y, N(0, I_d)) = \frac{1}{N} \sum_i \left\| \sigma \odot \alpha_i + \sigma^{-1} \odot (x_i - \mu) \right\|^2 \tag{7}$$

This is a special case of the affine transformation in Appendix A and minimal if $\sigma^2 = \mathrm{Var}[x_i]^{\frac{1}{2}} \mathrm{Var}[\alpha_i]^{-\frac{1}{2}}$ and $\mu = \bar{x}_i + \sigma^2 \bar{s}_i$; the same result from the solvable case in Section 2.1. This solution is very computationally inexpensive, and is hence the default in nutpie. Using Welford's algorithm to keep online estimates of the draw and score variances during sampling (thereby avoiding the need to explicity store scores), the mass matrix is set to a diagonal matrix with the i'th entry on the diagonal equal to $\mathrm{Var}[x_i]^{\frac{1}{2}} \mathrm{Var}[\alpha_i]^{-\frac{1}{2}}$.

Some theoretical results for normal posteriors

If the posterior is $N(\mu, \Sigma)$, then the minimizers μ^* and σ^* of $\hat{\mathcal{F}}$ converge to μ and $\exp\left(\frac{1}{2}\log\operatorname{diag}(\Sigma) - \frac{1}{2}\log\operatorname{diag}(\Sigma^{-1})\right)$, respectively. This is a direct consequence of $\operatorname{Cov}[x_i] \to \Sigma$ and $\operatorname{Cov}[\alpha_i] \to \Sigma^{-1}$.

$\hat{\mathcal{F}}$ converges to $\sum_i \lambda_i + \lambda_i^{-1}$, where λ_i are the generalized eigenvalues of Σ with respect to $\operatorname{diag}(\hat{\sigma}^2)$, so large and small eigenvalues are penalized. When we choose $\operatorname{diag}(\Sigma)$ as mass matrix, we effectively minimize $\sum_i \lambda_i$, and only penalize large eigenvalues. If we choose $\operatorname{diag}(\mathbb{E}(\alpha\alpha^T))$ (as proposed, for instance, in Tran and Kleppe (2024)) we effectively minimize $\sum \lambda_i^{-1}$ and only penalize small eigenvalues. But based on theoretical results for multivarite normal posteriors in Langmore et al. (2020), we know that both large and small eigenvalues make HMC less efficient. To this effect, we

We can use the result in (todo ref) to evaluate the different diagonal mass matrix choices on various gaussian posteriors, with different numbers of observations. Figure todo shows the resulting condition numbers of the posterior as seen by the sampler in the transformed space.

2.4.2 Full mass matrix

We choose $f_{A,\mu}(y) = Ay + \mu$. This corresponds to a mass matrix $M = (AA^T)^{-1}$. Because as we will see $\hat{\mathcal{F}}$ only depends on AA^T and μ, we can restrict A to be symmetric positive definite. We get

$$\hat{\mathcal{F}}[f] = \frac{1}{N}\sum \left\| A^T s_i + A^{-1}(x_i - \mu) \right\|^2 \tag{8}$$

which is minimized when $AA^T \operatorname{Cov}[x_i] AA^T = \operatorname{Cov}[\alpha_i]$ (proof in Appendix A), and as such corresponds again to our earlier derivation in Section 2.1. If the two covariance matrices are full rank, we get a unique minimum at the geometric mean of $\operatorname{Cov}[x_i]$ and $\operatorname{Cov}[s_i]$.

2.4.3 Diagonal plus low-rank

If the number of dimensions is larger than the number of available draws, we can add regularization terms. We must regularize so that the draws and scores share a common scale, otherwise the two sources of information might have dramatically different weights in the calculation of Σ. And to avoid $O(n^3)$ computation costs, we can project draws and scores into the span of x_i and α_i and compute the regularized mean in this subspace. We can also avoid $O(n^2)$ storage, and still do all operations we need for HMC, if we only store the eigenvectors and eigenvalues. To further reduce computational cost, we can ignore eigenvalues that are close to one with a cutoff parameter c. The algorithm is as follows:

LOW-RANK-ADAPT(D, G, c, γ):

$U^D \leftarrow \text{SVD}(D), U^G \leftarrow \text{SVD}(G)$

$S \leftarrow [U^D\ U^G]$ // Combine bases

$Q, \_ \leftarrow \text{QR-THIN}(S)$ // Get jointly-spanned orthonormal basis

$P^D \leftarrow Q^T D, P^G \leftarrow Q^T G$ // Project draws, grads onto shared subspace

$C^D \leftarrow P^D (P^D)^T + \gamma I$ // Get empirical covariance matrices, regularize

$C^G \leftarrow P^D (P^D)^T + \gamma I$

$\Sigma \leftarrow \text{SPDM}(C^D, C^G)$ // Solve $\Sigma C^G \Sigma = C^D$ for Σ

$U \Lambda U^{-1} \leftarrow \text{EIGENDECOMPOSE}(\Sigma)$ // Extract eigenvalues to subset

$U_c \leftarrow \{U_i : i \in \{i : \lambda_i \geq c \text{ or } \leq \frac{1}{c}\}\}$

$\Lambda_c \leftarrow \{\lambda_i : i \in \{i : \lambda_i \geq c \text{ or } \leq \frac{1}{c}\}\}$

$M \leftarrow Q U_c (\Lambda_c - 1) U_c^T + I$

return M

By subtracting the identity matrix from the diagonal matrix of eigenvalues, we obtain a matrix whose eigenvalues include all those of Σ which are sufficiently far from 1, with the remaining eigenvalues equal to 1. This is then a "diagonal plus low-rank" matrix.

3 Adaptation Schema

Whether we adapt a mass matrix using the posterior variance as Stan does, or if we use a bijection based on the Fisher divergence, we always have the same problem: in order to generate suitable posterior draws we need a good mass matrix (or bijection), but to estimate a good mass-matrix, we need posterior draws. There is a well known way out of this "chicken and egg" conundrum: we start sampling with an initial transformation, and collect a number of draws. Based on those draws, we estimate a better transformation, and repeat. This adaptation-window approach has long been used in the major implementations of HMC, and has remained largely unchanged for a number of years. PyMC, Numpyro, and Blackjax all use the same details as Stan, with at most minor modifications. There are, however, a couple of small changes that improve the efficiency of this schema significantly.

3.1 Choice of initial diffeomorphism

Stan's sampler begins its first adaptation using an identity mass matrix. The very first HMC trajectories seem to be overall much more reasonable if we use $M = \text{diag}(\alpha_0^T \alpha_0)$ instead. This also makes the initialization independent of variable scaling.

3.2 Accelerated Window-Based Adaptation

Stan and other samplers do not run vanilla HMC, but usually variants, most notably the No-U-Turn Sampler (NUTS) (Hoffman and Gelman 2011), where the length of the Hamiltonian trajectory is chosen dynamically. This can make it very costly to generate draws if the mass matrix is not well adapted, because in those cases we often use a large number of HMC steps for each draw (typically up to 1000). Thus very early on during sampling, we have a large incentive to use available information about the posterior as quickly as possible, to avoid these long trajectories. By default, Stan starts adaptation with a step-size adaptation window of 75 draws, where the mass matrix is untouched. This is followed by a mass matrix adaptation window consisting of a series of "memoryless" intervals of increasing length, the first of which (25 draws) still uses the initial mass matrix for sampling. These 100 draws before the first mass matrix change can constitute a sizable percentage of the total sampling time.

Intuition might suggest that we could just use a tailing window and update at each step based on the previous k draws. However, this is computationally inefficient (unless the logp function is very expensive), and is not easily implemented as a streaming estimator (see below for more details). Using several overlapping estimation windows, though, we can compromise between optimal information usage and computational cost, while still using streaming estimators.

4 Implementation in nutpie

nutpie implements:

- New adaptation schema
- Diagonal mass matrix adaptation based on the Fisher divergence
- Low rank mass matrix adaptation as described here
- Explicit transformation with normalizing flows

The core algorithms are implemented in rust, which all array operations abstracted away, to allow users of the rust API to provide GPU implementations.

It can take PyMC or Stan models.

Stan is used through bridgestan, which compiles C libraries that nutpie can load dynamically to call the logp function gradient with little overhead.

pymc models can be sampled either through the numba backend of pymc, which also allows evaluating the density and its gradient with little overhead. Alternatively, it can use the pymc jax backend. This incures a higher per-call overhead, but allows evaluating the density on the GPU, which can significantly speed up sampling for larger models.

nutpie returns sampling traces as arviz datasets, to allow easy posterior analysis and convergence checks.

5 Experimental evaluation of nutpie

We run nutpie and cmdstan on posteriordb to compare performance in terms of effective sample size per gradient evaluation and in terms of effective sample size per time…

LEAPFROG$(\theta,\ \rho,\ L,\ h)$:
$\theta^0 \leftarrow \theta, \rho^0 \leftarrow \rho$
for i from 0 **to** L:
    $\rho^{(i+\frac{1}{2})} \leftarrow \rho^{(i)} - \frac{h}{2}\nabla U\big(\theta^{(i)}\big)$ // half-step momentum
    $\theta^{(i+1)} \leftarrow \theta^{(i)} + h\rho^{(i+\frac{1}{2})}$ // full-step position
    $\rho^{(i+1)} \leftarrow \rho^{(i+\frac{1}{2})} - \frac{h}{2}\nabla U\big(\theta^{(i+1)}\big)$ // half-step momentum
return $\big(\theta^{(L)}, \rho^{(L)}\big)$

NUTS$(\theta,\ h,\ \varepsilon,\ M)$:
$\rho \sim N(0, I_{d\times d})$ // refresh momentum
$B \sim \text{Unif}(\{0,1\}^M)$ // resample Bernoulli process
$(a, b, \_) \leftarrow$ ORBIT-SELECT $(\theta, \rho, B, \varepsilon)$
$(\theta^*, \_, \_) \leftarrow$ INDEX-SELECT $(\theta, \rho, a, b, \varepsilon)$
return θ^*

ORBIT-SELECT$(\theta,\ \rho,\ B,\ \varepsilon)$:
$a, b \leftarrow 0$
for i from 0 **to** len(B):
    $\tilde{a} \leftarrow a + (-1)^{B_i}2^{i-1}, \tilde{b} \leftarrow b + (-1)^{B_i}2^{i-1}$
    $\mathbb{I}_{\text{U-Turn}} \leftarrow$ U-Turn $(a, b, \theta, \rho, \varepsilon)$
    $\mathbb{I}_{\text{Sub-U-Turn}} \leftarrow$ Sub-U-Turn $\big(\tilde{a}, \tilde{b}, \theta, \rho, \varepsilon\big)$
    if $\max(\mathbb{I}_{\text{U-Turn}}, \mathbb{I}_{\text{Sub-U-Turn}}) = 0$:
        $a \leftarrow \min(a, \tilde{a}), b \leftarrow \max\big(b, \tilde{b}\big)$
    else:
        break
return $a, b, \nabla H$

INDEX-SELECT$(\theta,\ \rho,\ a,\ b,\ \varepsilon)$:

$a, b \leftarrow 0$

for i from 0 **to** $\text{len}(B)$:

$\tilde{a} \leftarrow a + (-1)^{B_i} 2^{i-1}, \tilde{b} \leftarrow b + (-1)^{B_i} 2^{i-1}$

$\mathbb{I}_{\text{U-Turn}} \leftarrow$ U-Turn $(a, b, \theta, \rho, \varepsilon)$

$\mathbb{I}_{\text{Sub-U-Turn}} \leftarrow$ Sub-U-Turn $\left(\tilde{a}, \tilde{b}, \theta, \rho, \varepsilon\right)$

if $\max(\mathbb{I}_{\text{U-Turn}}, \mathbb{I}_{\text{Sub-U-Turn}}) = 0$:

$a \leftarrow \min(a, \tilde{a}), b \leftarrow \max\left(b, \tilde{b}\right)$

else:

break

return $a, b, \nabla H$

Bibliography

[1] B. Carpenter *et al.*, "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, vol. 76, no. 1, pp. 1–32, 2017, doi: 10.18637/jss.v076.i01.

[2] R. M. Neal, "MCMC using Hamiltonian dynamics." Accessed: Nov. 26, 2024. [Online]. Available: http://arxiv.org/abs/1206.1901

[3] J. H. Tran and T. S. Kleppe, "Tuning diagonal scale matrices for HMC." Accessed: Nov. 26, 2024. [Online]. Available: http://arxiv.org/abs/2403.07495

[4] I. Langmore, M. Dikovsky, S. Geraedts, P. Norgaard, and R. Von Behren, "A Condition Number for Hamiltonian Monte Carlo." Accessed: Oct. 16, 2022. [Online]. Available: http://arxiv.org/abs/1905.09813

[5] M. D. Hoffman and A. Gelman, "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." [Online]. Available: https://arxiv.org/abs/1111.4246

A Minimize Fisher divergence for Affine Transformations

Here we prove that $\hat{F}$ for $F(y) = Ay + \mu$ is minimal when $\Sigma\,\mathrm{Cov}[\alpha]\Sigma = \mathrm{Cov}[x]$ and $\mu = \bar{x} + \Sigma\bar{\alpha}$, where $\Sigma = AA^T$:

Let G be the matrix of scores, with α_i as the ith column, and similarly let X be the draws matrix, consisting of x_i as the i'th column, and $\Sigma = AA^T$. The Fisher divergence between some p and $N(0, I_d)$ is

$$E_p[\|\nabla \log p(x) + X\|^2] \tag{9}$$

and as such the estimated divergence is

$$\hat{F} = \frac{1}{N}\|G + X\|^2 \tag{10}$$

Now, for some transformed $y = A^{-1}(x - \mu)$, we have

$$\hat{F}_y = \frac{1}{N}\left\|A^T G + Y\right\|^2 = \frac{1}{N}\left\|A^T G + A^{-1}(X - \mu 1^T)\right\|_F^2 \tag{11}$$

Differentiating with respect to μ, we have:

$$\frac{d\hat{F}}{d\mu} = -\frac{2}{N}\,\mathrm{tr}\Big[\mathbf{1}^T\left(A^T G + A^{-1}(X - \mu\mathbf{1}^T)\right)^T A^{-1}\Big] \tag{12}$$

Setting this to zero,

$$\mathbf{1}^T\left(A^T G + A^{-1}(X - \mu\mathbf{1}^T)\right)^T A^{-1} = 0 \tag{13}$$

It follows that

$$\mu^* = \bar{x} + \Sigma\bar{\alpha} \tag{14}$$

Differentiating with respect to A:

$$d\hat{F} = \frac{2}{N}\,\mathrm{tr}\Big[\left(A^T G + A^{-1}(X - \mu\mathbf{1}^T)\right)^T\left(dA^T G - A^{-1}dAA^{-1}(X - \mu\mathbf{1}^T)\right)\Big] \tag{15}$$

Plugging in the result for μ^* and using the cyclic- and transpose-invariance properties of the trace gives us

$$\begin{aligned} d\hat{F} = &\frac{2}{N}\,\mathrm{tr}\big[\left(A^T G + A^{-1}\left(\tilde{X} - \Sigma\bar{\alpha}\mathbf{1}^T\right)\right)G^T dA\big] \\ &+\frac{2}{N}\,\mathrm{tr}\Big[A^{-1}\left(\Sigma\bar{\alpha}\mathbf{1}^T - \tilde{X}\right)\left(A^T G + A^{-1}\left(\tilde{X} - \Sigma\bar{\alpha}\mathbf{1}^T\right)\right)^T A^{-1}dA\Big] \end{aligned} \tag{16}$$

where $\tilde{X} = X - \bar{x}\mathbf{1}^T$, the matrix with centered x_i as the columns. This is zero for all dA iff

$$\begin{aligned} 0 &= \left(A^T G + A^{-1}\left(\tilde{X} - \Sigma\bar{\alpha}\mathbf{1}^T\right)\right)G^T + A^{-1}\left(\Sigma\bar{\alpha}\mathbf{1}^T - \tilde{X}\right)\left(A^T G + A^{-1}\left(\tilde{X} - \Sigma\bar{\alpha}\mathbf{1}^T\right)\right)^T A^{-1} \\ &= A^T\tilde{G}G^T + A^{-1}\tilde{X}G^T + \left(A^T\bar{\alpha}\mathbf{1}^T - A^{-1}\tilde{X}\right)\left(\tilde{G}^T + \tilde{X}^T\Sigma^{-1}\right), \end{aligned} \tag{17}$$

Where similarly $\tilde{G} = G - \bar{\alpha}\mathbf{1}^T$. Because $\mathbf{1}^T\tilde{X}^T = \mathbf{1}^T\tilde{G}^T = 0$, this expands to

$$\begin{aligned} 0 &= A^T\tilde{G}G^T + A^{-1}\tilde{X}G^T - A^{-1}\tilde{X}\tilde{G}^T - A^{-1}\tilde{X}\tilde{X}^T\Sigma^{-1} \\ &= \left(\tilde{G} + \Sigma^{-1}\tilde{X}\right)G^T - \Sigma^{-1}\tilde{X}\tilde{G}^T - \Sigma^{-1}\tilde{X}\tilde{X}^T\Sigma^{-1} \end{aligned} \tag{18}$$

$$\begin{aligned} &= \tilde{G}G^T + \Sigma^{-1}\tilde{X}\mathbf{1}\bar{\alpha}^T - \Sigma^{-1}\tilde{X}\tilde{X}^T\Sigma^{-1} \\ &= \tilde{G}\tilde{G}^T - \Sigma^{-1}\tilde{X}\tilde{X}^T\Sigma^{-1} \end{aligned} \tag{19}$$

Code: https://github.com/pymc-devs/nutpie