We would like to thank all referees for their appreciation of our results and the useful feedback. Below is our reply.

Reviewer 1: The regularized version of the FR problem is a geodesically convex optimization problem over the feasible set $\mathbb{S}^n_{++}$. However, the regularized problem has two drawbacks: *(i)* its objective function is not g-Lipschitz continuous over $\mathbb{S}^n_{++}$ because of the term $\langle S, \Sigma^{-1}\rangle$, *(ii)* $\mathbb{S}^n_{++}$ has infinite diameter. Due to these obstacles, the algorithms in [42, 41] cannot be used to solve the regularized problem. To the best of our knowledge there is no algorithm in the literature which can be readily applied to solve the regularized problem with convergence guarantee. By constraining the feasible set to $\mathcal{B}^{\rm FR}$, we can overcome these technical difficulties to establish the convergence guarantee in Theorem 2.7.

The KL divergence (confined to the subspace of Gaussian distributions) is not induced by any Riemannian metric. If we view problem (12) as a manifold optimization problem with the same Riemannian geometry as in Sect. 2, then (12) and its regularized version are both geodesically convex problems. However, problem (12) and its regularized version are convex in the usual Euclidean sense under the reparametrization $X = \Sigma^{-1}$, and it is more efficient to solve problem (12) by applying Theorem 3.2. Empirically, for dimension $d = 100$, on average solving the KL problem (12) using Theorem 3.2 takes < 0.1 seconds, while solving the FR problem (6) takes 1 second using Algorithm 1.

Reviewer 2: Your main suggestions for improvement focus around three aspects of the manuscript:

1. Connection between FR and KL: We apologize for not motivating thoroughly why we study both FR and KL. Ideally, we would like to use the FR metric since it is the unique metric that possesses the powerful invariance properties discussed in eqs. (4) and (5). These properties imply, amongst others, that the FR metric is invariant to the coordinate basis that frequently needs to be chosen arbitrarily in geometric problems. While failing to satisfy these desirable properties, the KL divergence constitutes an approximation to the FR metric (as discussed in footnote 1 of the appendix) that is computationally more tractable. We propose to elaborate on these connections in the introduction. To further illustrate the commonalities and differences between the two approaches, we also propose to replace Section 5.1 (which, as you correctly pointed out, does not add much insight) with a section that visualizes and compares the decision boundaries of the nominal QDA and those of FR and KL in the context of our application. In particular, we observe that our approaches lead to non-hyperbolic decision boundaries in general. We will also add a comparison of the wallclock times in Section 5.2. As we pointed out in our response to Rev. 1, solving the KL problem (12) using Theorem 3.2 takes < 0.1 seconds on average for dimension $d = 100$, while solving the FR problem (6) takes 1 second using Algorithm 1. Because (12) is non-convex, the gradient descent algorithm cannot guarantee to converge to global minimum of (12).

2. Further explanations for Sections 4+5: Thank you for pointing out the lack of explanation in the main paper regarding the ambiguity set used in this application. Appendix A of the manuscript argues that for a fixed sample size, estimating $\hat{\Sigma}_c$ is much harder than estimating $\hat{\mu}_c$. In our numerical experiments, we thus identify $\hat{\mu}_c$ with the sample average and only consider uncertainty in $\hat{\Sigma}_c$. We will add a discussion in the paper that summarizes our findings from the appendix and clarifies which ambuiguity set we use.

3. Intuition for the use of the ergodic geodesic average in Algorithm 1: Thank you for pointing out this omission. The ergodic geodesic average $\bar{\Sigma}_k$ is the sequence that has been proven to converge in [42], and we are not aware of any last-iterate convergence results under the similar conditions of problem (6). We propose to clarify this aspect in the camera-ready version of the paper.

Thank you also for your minor suggestions, which we plan to address in the revised version of the manuscript.

Reviewer 3: We apologize for the lack of rigor in our use of the term "computationally intractable". We meant to say that the problem is non-convex (since the L^2-Wasserstein manifold of Gaussian measures has a non-negative sectional curvature (see [36]), and the objective function is not geodesically convex on this manifold) and therefore *appears* to be computationally intractable, but we do not have a rigorous hardness result. We will fix this in the revised version. We also agree that the Fisher information matrix may be rank deficient if we have a degenerate Gaussian distribution, in which case the inner product on the tangent space would fail to be positive definite. Since we work with the set $\mathcal{M}$ of all *non-degenerate* Gaussian distributions (cf. line 34 of the paper), however, the Fisher information matrix will always be positive definite. We will highlight this in the revised version of the paper. As for footnote 2, thank you for pointing out that the circle has zero intrinsic curvature; we will update the paper accordingly. In line 173, we will replace the statements "closed form" and "highly efficient" with the appropriate complexity estimates. The proof of Theorem 9 in [42] involves bounding the gradient via using the Lipschitz assumption. For this argument to be valid, the Lipschitz assumption needs to hold on an open subset $\mathcal{Y} \subseteq \mathcal{M}$ containing $\mathcal{B}^{\rm FR}$. We simplified the proof a little bit by directly bounding the gradient on $\mathcal{B}^{\rm FR}$ which can be done for our problem. Regarding linear convergence rate: under the (minor) assumption that $S \succ 0$, we can show that the objective function is strongly g-convex and g-smooth over the ball $\mathcal{B}^{\rm FR}$ (the explicit constants can be computed from $\lambda_{\min}(S)$, $\lambda_{\max}(S)$ and the bounds in Lemma C.1). Theorem 15 in [42] applies, and Algorithm 1 (with an appropriately modified stepsize) converges linearly for the sequence Σ_k. We will add this result in the revised version. Empirically, we observe the linear convergence rate even when S is singular. Thank you very much for your suggestion!