# Barycenters vs. Model Averages in Gaussian Models

## Contents



## 1 Problem Description

It is of importance in machine learning to be able to combine models, be it as a means of reducing uncertainty in predictions or hedging risk during prediction time. The machine learning community has been exploring this topic in various forms, from ensembling techniques ([LPB17]) to stochastic optimal transport ([CCS18]). A recent paper, [MF17], explores the idea of averaging Gaussian Process models using Wasserstein Barycenters, and show that it produces clearer and more representative results than the "naive average", which is equivalent to calculating the average and standard deviation of the mean of these gaussian processes. They propose a tractable fixed point iteration method for calculating the barycenter distribution, and display experimental results. A question to ask is whether this is required for Gaussian Processes at all, because averaging finite dimensional representations of these is equivalent to adding up Gaussian measures. Identifying similarities and differences between these two approaches can help the community understand when the $O(n^3)$ complexity of calculating the barycenter is a requirement, or when a simple sum of gaussian moments could be used.

The probability of a predicted function $f$ from a model within a set of models $M$ of

size $N$ is as follows:

$$p(f) = \int p(f|M)p(M)dM \tag{1}$$

which follows from marginalization. This method of reasoning about the "average" of model predictions will be termed *Model Averaging*. For the case of gaussian process models, the integral in 1 is tractable, as the sum becomes discrete and gaussians are closed under addition:

$$p(f) = \sum_{i=1}^{N} \xi_i p(f|M_i) \tag{2}$$

where $\xi$ are the discrete probabilities of each model (assuming $p(M)$ is a discrete measure). For gaussian finite dimensional distributions (the stochastic processes evaluated at discrete input points) $X_i$, $p(f)$ has the following moments:

$$\mathbb{E}[X^*] = \sum_{i=1}^{N} \xi_i \mathbb{E}[X_i], \qquad \mathbb{E}[X^* X^{*T}] = \sum_{i=1}^{N} \xi_i \mathbb{E}[X_i X_i^T],$$

Where $X^*$ is the random variable denoting $p(f)$ in 2. It is not always the case, however, that 1 is analytically tractable, and in such cases the integral and moments of the distribution can be approximated using Monte Carlo.

The barycenter $\mu^*$ of $N$ gaussian distributions $\mu_i$ is defined as follows:

$$\mu^* = \inf_{\mu \in P_2(H)} \sum_{i=1}^{N} \xi_i W_2^2(\mu_i, \mu)$$

where $W_2^2(\mu_1, \mu_2)$ denotes the 2-Wasserstein metric between gaussian measures $\mu_1$, $\mu_2$ [MF17].

It is unclear whether, for their use in GPs, barycenters provide a reduced uncertainty ensemble of models compared to standard model averaging. The text explores this idea, and states certain scenarios where these two methods of averaging are equal, and

where they diverge.

# 2 Theoretical Observations

## 2.1 Equivalent Covariance Matrices

For the finite dimensional distribution representation of Gaussian Processes, it can be shown that the barycenter distribution is a Gaussian distribution with mean $\bar{m}$ and covariance matrix $\bar{K}$ denoted as follows [MF17]:

$$\bar{m} = \sum_{i=1}^{N} \xi_i m_i, \qquad \bar{K} = \sum_{i=1}^{N} \xi_i (\bar{K}^{\frac{1}{2}} K_i \bar{K}^{\frac{1}{2}}),$$

Assume there is some data $\mathbf{y}$ and some time points we would like to predict $t_*$ with function values $y_*$. This can be modelled with a GP:

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \tag{3}$$

Conditioning this GP on a training set $\mathbf{y}$ and evaluating on the finite set $T^*$ produces a finite dimensional gaussian distribution with moments [Ebd15]:

$$\mathbb{E}[\bar{y}_*] = K_* K^{-1} \mathbf{y}, \qquad \hat{K} = var(\bar{y}_*) = K_{**} - K_* K^{-1} K_*^T$$

Let us consider the case where we are averaging GPs with equivalent kernel functions $k(t, t')$ and means $\mu_i$. Assume that our prior distribution over these models is uniform, so $\xi_i = 1/N$. The model average $p(f)$ can be calculated from 2 and the fact that gaussians are closed under addition:

$$p(f) \sim \mathcal{N}(\frac{1}{2}\mu_1 + \frac{1}{2}\mu_2, \frac{1}{2}K_1 + \frac{1}{2}K_2) \tag{4}$$

Where $K_i$ denotes the covariance matrix that is the gram matrix produced from a GP's respective kernel function $k(t, t')$. According to our assumption that both GPs contain equivalent kernel functions, their covariance matrices $K_i$ are equivalent, so 4 reduces to:

$$p(f) \sim \mathcal{N}(\frac{1}{2}\mu_1 + \frac{1}{2}\mu_2, K) \tag{5}$$

where $K = K_1 = K_2$. Now, since we currently care about averaging the predictions of GPs that have been conditioned on data (they are not as informative otherwise), we can note that the posterior kernels of these two GPs are also equivalent, which follows from the equivalence of their kernel functions and the equivalent formulations for $\hat{K}$, the posterior covariance matrix. Let us denote this common posterior covariance matrix $\hat{K}$, and consider the following proposition:

**Proposition 1.** The barycenter of gaussian processes with equivalent kernel functions $k(t, t')$ is equivalent to their model average.

*Proof.* In order to show this is true, we have to show that both the moments of both distributions (the barycenter and the model average) are equivalent, since gaussian measures are entirely defined by their first two moments. We start with the means which follow trivially from 2 and 3:

$$\mathbb{E}[p(f)] = \sum \xi_i \mathbb{E}[p(f|M)] = \sum \xi_i \mu_i = \bar{m} \tag{6}$$

We now work to show that $\bar{K} = \hat{K}$. This can be shown by proving that $\hat{K}$ is a solution to the fixed point iteration equation 2. This equation is shown to be convex in [MF17], so finding a solution ensures that it is optimal:

$$\bar{K} = \sum_{i=1}^{N} \xi_i (\bar{K}^{\frac{1}{2}} K_i \bar{K}^{\frac{1}{2}}) \tag{7}$$

Substitute $\hat{K}$ into the right hand side:

$$\sum_{i=1}^{N} \xi_i (\hat{K}^{\frac{1}{2}} K_i \hat{K}^{\frac{1}{2}}) \tag{8}$$

but keeping in mind that, since all gaussian processes possess the same kernel function, $\hat{K} = K_i$, so:

$$= \sum_{i=1}^{N} \xi_i (\hat{K}^{\frac{1}{2}} \hat{K} \hat{K}^{\frac{1}{2}})^(\frac{1}{2}) \tag{9}$$

We have to show that 9 equals $\hat{K}$ for the implicit equation to be satisfied. Keeping in mind that covariance matrices are symmetric, hence positive semidefinite, diagonalizable, and possess orthogonal eigenvectors, we apply the diagonal decomposition to each matrix:

$$\hat{K} = U \Lambda U^T \tag{10}$$

which leads to

$$= \sum_{i=1}^{N} \xi_i ((U \Lambda U^T)^{\frac{1}{2}} (U \Lambda U^T)(U \Lambda U^T)^{\frac{1}{2}}). \tag{11}$$

Recall that, for orthogonal matrices $U$, $UU^T = I$ and

$$(U \Lambda^{\frac{1}{2}} U^T)(U \Lambda^{\frac{1}{2}} U^T) = U \Lambda U^T \tag{12}$$

so, $(U \Lambda^{\frac{1}{2}} U^T) = (U \Lambda U^T)^{\frac{1}{2}}$. Plugging this into 9:

$$\begin{aligned}
&= \sum_{i=1}^{N} \xi_i ((U \Lambda^{\frac{1}{2}} U^T)(U \Lambda U^T)(U \Lambda^{\frac{1}{2}} U^T))^{\frac{1}{2}} \\
&= \sum_{i=1}^{N} \xi_i (U \Lambda^{\frac{1}{2}} \Lambda \Lambda^{\frac{1}{2}} U^T)^{\frac{1}{2}} \\
&= \sum_{i=1}^{N} \xi_i (U \Lambda^2 U^T)^{\frac{1}{2}} \\
&= U \Lambda U^T \\
&= \hat{K}
\end{aligned}$$

Therefore, $\hat{K}$ is the unique solution to the fixed point iteration equation and we have shown that the gaussian distributions created by model averaging and calculating the barycenter of gaussian processes with prior probabilities $x_i$ are equivalent given that they possess the same kernel function $k(t, t')$. □

This result entails that conclusions achieved in [MF17] could have been achieved through simple model averaging, as the paper used the same kernels for fitting the gaussian processes, but different samples of data.

## 2.2 Covariance Matrices differing by a constant

For the case of covariance matrices differing by a constant, we analyze the following finite dimensional GPs:

$$\mathbf{y} \sim \mathcal{N}\big(0, \Sigma\big), \mathbf{y} \sim \mathcal{N}\big(0, \alpha\Sigma\big) \tag{13}$$

where $\alpha$ is a positive constant. Analyzing these two gaussians for now, we can attempt to derive what a solution to the fixed point iteration barycenter equation could be, suppose it is of the form $\beta\Sigma$, $\beta$ a positive constant. Then from 7 and assuming a uniform average of distributions ($\xi = 1/N$) we expand:

$$\beta\Sigma = \frac{1}{2}((\beta\Sigma)^{\frac{1}{2}}\Sigma(\beta\Sigma)^{\frac{1}{2}})^{\frac{1}{2}} + \frac{1}{2}((\beta\Sigma)^{\frac{1}{2}}\alpha\Sigma(\beta\Sigma)^{\frac{1}{2}})^{\frac{1}{2}} \tag{14}$$

And, taking into account that $\Sigma$ is a positive definite matrix, we use orthogonality to simplify:

$$\beta\Sigma = \frac{1}{2}((\beta\Sigma)^{\frac{1}{2}}\Sigma(\beta\Sigma)^{\frac{1}{2}})^{\frac{1}{2}} + \frac{1}{2}((\beta\Sigma)^{\frac{1}{2}}\alpha\Sigma(\beta\Sigma)^{\frac{1}{2}})^{\frac{1}{2}} \tag{15}$$

$$\beta\Sigma = \frac{1}{2}\beta^{\frac{1}{2}}((\Sigma)^{\frac{1}{2}}\Sigma(\Sigma)^{\frac{1}{2}})^{\frac{1}{2}} + \frac{1}{2}\beta^{\frac{1}{2}}\alpha^{\frac{1}{2}}((\Sigma)^{\frac{1}{2}}\Sigma(\Sigma)^{\frac{1}{2}})^{\frac{1}{2}} \tag{16}$$

$$\beta\Sigma = ((\Sigma)^{\frac{1}{2}}\Sigma(\Sigma)^{\frac{1}{2}})^{\frac{1}{2}} * (\frac{1}{2}\beta^{\frac{1}{2}} + \frac{1}{2}\beta^{\frac{1}{2}}\alpha^{\frac{1}{2}}) \tag{17}$$

$$\frac{\beta}{(\frac{1}{2}\beta^{\frac{1}{2}} + \frac{1}{2}\beta^{\frac{1}{2}}\alpha^{\frac{1}{2}})}\Sigma = ((\Sigma)^{\frac{1}{2}}\Sigma(\Sigma)^{\frac{1}{2}})^{\frac{1}{2}} \tag{18}$$

$$\frac{\beta}{(\frac{1}{2}\beta^{\frac{1}{2}} + \frac{1}{2}\beta^{\frac{1}{2}}\alpha^{\frac{1}{2}})} = 1 \tag{19}$$

$$\beta^{\frac{1}{2}}(\beta^{\frac{1}{2}} - (\frac{1}{2} + \frac{1}{2}\alpha^{\frac{1}{2}})) = 0 \tag{20}$$

$$\tag{21}$$

simplifying taking into account the relation shown earlier, $K = (K^{\frac{1}{2}}KK^{\frac{1}{2}})^{frac12}$ for positive definite matrices. Ignoring the trivial solution $\beta^{frac12} = 0$, we get that:

$$\beta^{\frac{1}{2}} = (\frac{1}{2} + \frac{1}{2}\alpha^{\frac{1}{2}}) \tag{22}$$

And if we do the same procedure with the euclidean average, calculating $\beta_{EU}$ for $\Sigma_{EU}$ (also called the euclidean barycenter):

$$\beta_{EU}\Sigma_{EU} = \frac{1}{2}\Sigma_{EU} + \frac{1}{2}\alpha\Sigma_{EU} \tag{23}$$

$$\beta_{EU} = \frac{1}{2} + \frac{1}{2}\alpha \tag{24}$$

Now, if we analyze equations , and , we see that the first grows sublinearly with $\alpha$, while the second grows linearly with $\alpha$. This makes sense, and it shows that, for gaussians whose covariance matrices differ by a constant, the barycenter always has lower variance. The two $\beta$'s are equal when $\alpha = 1$, which is the first case we analyzed, in all other cases, $(\frac{1}{2} + \frac{1}{2}\alpha^{\frac{1}{2}}) < \frac{1}{2} + \frac{1}{2}\alpha$.

## 2.3 Entropy

Now, we can formulate the same problem using the notion of entropy, and finding which type of barycenter (euclidean or wassserstein) will yield averages with less entropy (more certainty). The entropy of a gaussian distribution of dimension $k$ is:

$$H = \frac{k}{2} + \frac{k}{2}log(2\pi) + \frac{1}{2}log|\Sigma| \tag{25}$$

Now we can analyze the entropy of the euclidean and wasserstein barycenters, denoted $H_{EU}$ and $H_{OT}$ respectively. Consider the simple case of computing the barycenter of two normal distributions of dimension $k$ with $\xi_i = \frac{1}{2}$:

$$H_{EU} = \frac{k}{2} + \frac{k}{2}log(2\pi) + \frac{1}{2}log|K_{EU}| \tag{26}$$

$$H_{OT} = \frac{k}{2} + \frac{k}{2}log(2\pi) + \frac{1}{2}log|K_{OT}| \tag{27}$$

$$\tag{28}$$

The only terms that differ at the covariance matrix determinants, so let's analyze those.

$$\begin{aligned}
|K_{OT}| &= |\frac{1}{2}(K_{OT}^{\frac{1}{2}}K_1K_{OT}^{\frac{1}{2}})^{\frac{1}{2}} + \frac{1}{2}(K_{OT}^{\frac{1}{2}}K_1K_{OT}^{\frac{1}{2}})^{\frac{1}{2}}| \\
|K_{OT}| &\geq |\frac{1}{2}(K_{OT}^{\frac{1}{2}}K_1K_{OT}^{\frac{1}{2}})^{\frac{1}{2}}| + |\frac{1}{2}(K_{OT}^{\frac{1}{2}}K_1K_{OT}^{\frac{1}{2}})^{\frac{1}{2}}| \\
|K_{OT}| &\geq \frac{1}{2}^k |K_{OT}|^{\frac{1}{2}}|K_1|^{\frac{1}{2}} + \frac{1}{2}^k |K_{OT}|^{\frac{1}{2}}|K_2|^{\frac{1}{2}} \\
(|K_{OT}|^{\frac{1}{2}})(|K_{OT}|^{\frac{1}{2}} - (\frac{1}{2}^k |K_1|^{\frac{1}{2}} + \frac{1}{2}^k |K_2|^{\frac{1}{2}})) &\geq 0 \\
(|K_{OT}|^{\frac{1}{2}} - (\frac{1}{2}^k |K_1|^{\frac{1}{2}} + \frac{1}{2}^k |K_2|^{\frac{1}{2}})) &\geq 0
\end{aligned}$$

Using the fact that $K_{OT}$ is positive definite and $|A + B| \geq |A| + |B|$ for matrices $A$ and $B$. Now if we analyze the euclidean barycenter covariance matrix:

$$|K_{EU}| = |\frac{1}{2}K_1 + \frac{1}{2}K_2|$$
$$|K_{EU}| \geq \frac{1}{2}^k |K_1| + \frac{1}{2}^k |K_2|$$

It can also be shown through KKT optimization that both matrix determinants can be upper bounded by $|K_j|$ where $K$

**TODO 1.**

- Loosen the restriction a little bit, and can start looking at covariance matrices that are simultaneously diagonalizable (same eigenvectors, but different eigenvalues)
- Start looking at the entropy of $K$, and how that changes. This has been started, and calculating the entropy actually gives you a lower bound on the $K$ derived from both methods. This is yet to be typed into Latex.
- Properly calculate the computational complexity of barycentering vs calculating the model average
- Comparing with the case where kernels produce covariance functions with the same eigenvectors, but potentially different eigenvalues and what that means
- Is there any case where the variance of the barycenter is larger?
- Any proofs for non-gaussian models?

# References

[CCS18] Sebastian Claici, Edward Chien, and Justin Solomon. "Stochastic Wasserstein Barycenters". In: *CoRR* abs/1802.05757 (2018). arXiv: 1802.05757. URL: http://arxiv.org/abs/1802.05757 (cit. on p. 1).

[Ebd15] Mark Ebden. *Gaussian Processes: A Quick Introduction*. 2015. eprint: arXiv:1505.02965 (cit. on p. 3).

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 6402–6413. URL: http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf (cit. on p. 1).

[MF17] Anton Mallasto and Aasa Feragen. "Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 5660–5670. URL: http://papers.nips.cc/paper/7149-learning-from-uncertain-curves-the-2-wasserstein-metric-for-gaussian-processes.pdf (cit. on pp. 1–4, 6).