Normalization in Gaussian Processes

Kaiwen Wu
University of Pennsylvania
kaiwenwu@seas.upenn.edu

December 2021

Abstract

We present the invariance property of Gaussian processes under linear transformations on the training labels. As a result, we show how to unnormalize the predictive distribution if the GP is trained on normalized labels.

1 Introduction

```
model = GP(
    train_x,
    train_y,
)
model.train()

mvn_func = model.predict_func(test_x)
mvn_grad = model.predict_grad(test_x)
```

(a) Unnormalized GP training

```
model = GP(
    train_x,
    (train_y - train_y.mean()) / train_y.std(),
)
model.train()

mvn_func = model.predict_func(test_x)
mvn_grad = model.predict_grad(test_x)
```

(b) Normalized GP training

Figure 1: Pseudo codes of unnormalized/normalized GP training.

Figure 1 shows the pseudo codes of two Gaussian processes trained on the same data. The only difference between them is that Figure 1b normalizes the training label before fitting the model. In practice, such normalization is often used to (a) enhance the numerical stability and (b) scale the training objective to make it "easier" to optimize numerically (see Lemma 1). Our goal is to understand the relation between the predictive distributions `mvn_func` and `mvn_grad` with and without the normalization.

Let $\mathbf{y} \in \mathbb{R}^n$ be the training label. Let $\mathbf{f}^*|\mathbf{f}$ and $\nabla\mathbf{f}^*|\mathbf{f}$ be the predictive distributions for the function value and the gradient, *i.e.*, `mvn_func` and `mvn_grad`. Intuitively, if we apply an element-wise linear transformation on the training label

$$\mathbf{y} \mapsto a\mathbf{y} + b,$$

where $a, b \in \mathbb{R}$, then the predictive distributions should be changed by the same linear transformation

$$(\mathbf{f}^*|\mathbf{f}) \mapsto a\,(\mathbf{f}^*|\mathbf{f}) + b$$
$$(\nabla\mathbf{f}^*|\mathbf{f}) \mapsto a\,(\nabla\mathbf{f}^*|\mathbf{f})\,,$$

which will be proved in the next section. In this note, we refer to this property as invariance under linear transformation, as the two GPs are essentially the same up to a linear transformation.

2 Proof

For simplicity, let us consider a Gaussian process with a constant mean function $\mu(\mathbf{x}) = \mu \in \mathbb{R}$ and a RBF kernel $k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2l^2}\right)$, where $l \in \mathbb{R}$ is the length scale. Though, the following argument immediately applies to more general mean functions and kernels.

The prior distribution is

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\mu, \begin{bmatrix} \alpha K_{\mathbf{f},\mathbf{f}} + sI & \alpha K_{\mathbf{f},*} \\ \alpha K_{*,\mathbf{f}} & \alpha K_{*,*} \end{bmatrix}\right),$$

where $\alpha \in \mathbb{R}$ is the output scale, $s \in \mathbb{R}$ is the noise variance, and $K_{\cdot,\cdot}$ is the covariance matrix by applying the RBF kernel to the training/test data.

The negative log marginal likelihood used in training is

$$\begin{aligned} \ell(\mu, \alpha, s, l; \mathbf{y}) &= -\log p(\mathbf{y}) \\ &\propto \frac{1}{2}(\mathbf{y} - \mu)^\top \left(\alpha K_{\mathbf{f},\mathbf{f}} + sI\right)^{-1} (\mathbf{y} - \mu) + \frac{1}{2} \log\det\left(\alpha K_{\mathbf{f},\mathbf{f}} + sI\right), \end{aligned} \tag{1}$$

where we (intentionally) hide the length scale variable l on the right hand side, the reason for which will become clear later. In fact, the length scale l plays no role in the following argument. More generally, all parameters inside the kernel play no role in the following argument.

We start with charactering the minimizers of the negative log likelihood (1) after applying the linear transformation $\mathbf{y} \mapsto a\mathbf{y} + b$.

Lemma 1. *Suppose that* $(\mu^\star, \alpha^\star, s^\star, l^\star)$ *is a minimizer of the negative log likelihood* (1). *Consider an arbitrary element-wise linear transformation* $\mathbf{y} \mapsto a\mathbf{y} + b$, *where* $a, b \in \mathbb{R}$ *and* $a \neq 0$. *Then, we have*

$$\left(a\mu^\star + b, a^2\alpha^\star, a^2 s^\star, l^\star\right) \in \operatorname*{argmin}_{\mu,\alpha,s,l} \ell(\mu, \alpha, s, l; a\mathbf{y} + b).$$

Proof. Straightforward calculation gives

$$\begin{aligned} \ell(\mu, \alpha, s, l; a\mathbf{y} + b) &= \frac{1}{2}(a\mathbf{y} + b - \mu)^\top \left(\alpha K + sI\right)^{-1} (a\mathbf{y} + b - \mu) + \frac{1}{2}\log\det\left(\alpha K + sI\right) \\ &= \frac{1}{2}\left(\mathbf{y} - \frac{\mu - b}{a}\right)^\top \left(\frac{\alpha}{a^2}K + \frac{s}{a^2}I\right)^{-1} \left(\mathbf{y} - \frac{\mu - b}{a}\right) + \frac{1}{2}\log\det\left(\frac{\alpha}{a^2}K + \frac{s}{a^2}I\right) + n\log a \\ &= \ell\left(\frac{\mu - b}{a}, \frac{\alpha}{a^2}, \frac{s}{a^2}, l; \mathbf{y}\right) + n\log a. \end{aligned}$$

Notice that the last term $n \log a$ is a constant that does not affect the minimizer. Since $(\mu^\star, \alpha^\star, s^\star, l^\star)$ minimizes $\ell(\mu, \alpha, s, l; \mathbf{y})$, it is immediate that $\left(a\mu^\star + b, a^2\alpha^\star, a^2 s^\star, l^\star\right)$ minimizes $\ell(\mu, \alpha, s, l; a\mathbf{y} + b)$. □

With Lemma 1, we are ready to prove the invariance property.

Proposition 1. *Assume the minimizer of the negative log likelihood* (1) *always exists and is unique. Suppose that the predictive distributions for the GP trained on the label* $\mathbf{y}$ *are*

$$\mathbf{f}^*|\mathbf{f} \sim \mathcal{N}\left(\mu_{\mathbf{f}^*|\mathbf{f}}, \Sigma_{\mathbf{f}^*|\mathbf{f}}\right), \nabla\mathbf{f}^*|\mathbf{f} \sim \mathcal{N}\left(\mu_{\nabla\mathbf{f}^*|\mathbf{f}}, \Sigma_{\nabla\mathbf{f}^*|\mathbf{f}}\right).$$

Then, the predictive distribution of the GP trained on the label $a\mathbf{y} + b$ *are*

$$\mathbf{f}^*|\mathbf{f} \sim \mathcal{N}\left(a\mu_{\mathbf{f}^*|\mathbf{f}} + b, a^2\Sigma_{\mathbf{f}^*|\mathbf{f}}\right), \nabla\mathbf{f}^*|\mathbf{f} \sim \mathcal{N}\left(a\mu_{\nabla\mathbf{f}^*|\mathbf{f}}, a^2\Sigma_{\nabla\mathbf{f}^*|\mathbf{f}}\right),$$

where $a, b \in \mathbb{R}$ *and* $a \neq 0$.

Proof. Let $(\mu^\star, \alpha^\star, s^\star, l^\star)$ be the (unique) minimizer of $\ell(\mu, \alpha, s, l; \mathbf{y})$. Then, for the GP trained using the label $\mathbf{y}$, we can write the predictive mean and covariance of $\mathbf{f}^*|\mathbf{f}$ as

$$\begin{aligned}
\mu_{\mathbf{f}^*|\mathbf{f}} &= \mu^\star + \alpha^\star K_{*,\mathbf{f}} \left(\alpha^\star K_{\mathbf{f},\mathbf{f}} + s^\star I\right)^{-1} (\mathbf{y} - \mu^\star) \\
\Sigma_{\mathbf{f}^*|\mathbf{f}} &= \alpha^\star K_{*,*} - (\alpha^\star K_{*,\mathbf{f}}) \left(\alpha^\star K_{\mathbf{f},\mathbf{f}} + s^\star I\right)^{-1} (\alpha^\star K_{\mathbf{f},*})
\end{aligned}$$

and write the predictive mean and covariance of $\nabla \mathbf{f}^*|\mathbf{f}$ as

$$\begin{aligned}
\mu_{\nabla\mathbf{f}^*|\mathbf{f}} &= \alpha^\star \nabla K_{*,\mathbf{f}} \left(\alpha^\star K_{\mathbf{f},\mathbf{f}} + s^\star I\right)^{-1} (\mathbf{y} - \mu^\star) \\
\Sigma_{\nabla\mathbf{f}^*|\mathbf{f}} &= \alpha^\star \nabla K_{*,*} \nabla^\top - (\alpha^\star \nabla K_{*,\mathbf{f}}) \left(\alpha^\star K_{\mathbf{f},\mathbf{f}} + s^\star I\right)^{-1} \left(\alpha^\star K_{\mathbf{f},*} \nabla^\top\right).
\end{aligned}$$

Applying Lemma 1, the (unique) minimizer of $\ell(\mu, \alpha, s, l; a\mathbf{y}+b)$ is $\left(a\mu^\star + b, a^2\alpha^\star, a^2 s^\star, l^\star\right)$. To get the predictive distributions for the new GP, replace $\mathbf{y}$ with $a\mathbf{y}+b$ and replace $(\mu^\star, \alpha^\star, s^\star, l^\star)$ with $\left(a\mu^\star + b, a^2\alpha^\star, a^2 s^\star, l^\star\right)$ in the above expressions. Straightforward calculation finishes the proof. □

Based on the proof of Proposition 1, the invariance property of GPs entirely rely on the hyperperameter optimization. In the case when the hyperparameter optimization is inexact, or some hyperparameters are fixed, the invariance property may not hold anymore.

Example 1. *Consider a Gaussian process with a fixed zero mean function, a RBF kernel, a trainable output scale and a trainable noise variance. Its predictive distribution is invariant under scaling* $\mathbf{y} \mapsto a\mathbf{y}$ *but is not invariant under shifting* $\mathbf{y} \mapsto \mathbf{y} + b$.

Finally, we point out that the uniqueness assumption in Proposition 1 is necessary. As Lemma 1 only says $\left(a\mu^\star + b, a^2\alpha^\star, a^2 s^\star, l^\star\right)$ is one of the minimizers of $\ell(\mu, \alpha, s, l; a\mathbf{y} + b)$, it becomes quite tricky if multiple global minimizers exist. In that case, the predictive distribution depends on the algorithmic choice in the hyperparameter optimization, and the GP prediction is not guaranteed to be invariant. Nevertheless, Proposition 1 should give strong confidence that Gaussian processes used in practice is invariant under label normalization.