A CONNECTION TO VARIATIONAL INFERENCE

A connection between Gaussian continuation and variational inference can be made. We start by defining our variational distribution as the Gaussian $q(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \lambda) = \mathcal{N}(\boldsymbol{\theta}, \lambda \boldsymbol{I})$ (equivalent to the Gaussian kernel k_λ defined in Section 2.1). Given a likelihood function $p(\mathcal{D}|\boldsymbol{\vartheta})$, and zero-mean prior $p(\boldsymbol{\vartheta}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, the ELBO is defined as

$$\text{ELBO}(\boldsymbol{\theta}, \lambda) = \mathbb{E}_{q(\boldsymbol{\vartheta}|\boldsymbol{\theta},\lambda)}\left[\log\left(p(\mathcal{D}|\boldsymbol{\vartheta})\right)\right] - D_{\text{KL}}(q(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \lambda) \,||\, p(\boldsymbol{\vartheta})), \tag{13}$$

where D_{KL} is the KL divergence. For the two Gaussian distributions $q(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \lambda)$ and $p(\boldsymbol{\vartheta})$, the KL divergence is

$$D_{\text{KL}}(q(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \lambda) \,||\, p(\boldsymbol{\vartheta})) = \frac{1}{2}\left(\log\left|\boldsymbol{\Sigma}(\lambda \boldsymbol{I})^{-1}\right| - m + \text{tr}\left(\lambda \boldsymbol{I}\boldsymbol{\Sigma}^{-1}\right) + \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right). \tag{14}$$

This may be further separated into terms that depend on $\boldsymbol{\theta}$ and λ,

$$\begin{aligned} D_{\text{KL}}(q(\boldsymbol{\vartheta}|\boldsymbol{\theta}, \lambda) \,||\, p(\boldsymbol{\vartheta})) &= \frac{1}{2}\left(\log|\boldsymbol{\Sigma}| - m\log(\lambda) - m + \lambda\text{tr}\left(\boldsymbol{\Sigma}^{-1}\right) + \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right) \\ &= \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \frac{1}{2}\left(\lambda\text{tr}\left(\boldsymbol{\Sigma}^{-1}\right) - m\log(\lambda)\right) + K, \end{aligned} \tag{15}$$

where K includes terms that are constant with respect to both $\boldsymbol{\theta}$ and λ.

We define our objective $f(\boldsymbol{\vartheta}) = -\log\left(p(\mathcal{D}|\boldsymbol{\vartheta})\right)$, so the expected value in equation 13 is $-\mathbb{E}_{q(\boldsymbol{\vartheta}|\boldsymbol{\theta},\lambda)}\left[f(\boldsymbol{\vartheta})\right]$. Then it follows that

$$-\mathbb{E}_{q(\boldsymbol{\vartheta}|\boldsymbol{\theta},\lambda)}\left[f(\boldsymbol{\vartheta})\right] = -\int_{\mathcal{M}} f(\boldsymbol{\vartheta})k_\lambda(\boldsymbol{\theta} - \boldsymbol{\vartheta})d\boldsymbol{\vartheta} = -\left[f \star k_\lambda\right](\boldsymbol{\theta}). \tag{16}$$

The ELBO may then be written as

$$\text{ELBO}(\boldsymbol{\theta}, \lambda) = -\left[f \star k_\lambda\right](\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - \frac{1}{2}\left(\lambda\text{tr}\left(\boldsymbol{\Sigma}^{-1}\right) - m\log(\lambda)\right) - K. \tag{17}$$

Maximizing this ELBO with respect to $\boldsymbol{\theta}$ here means finding the mean of the variational distribution q with fixed covariance $\lambda\boldsymbol{I}$. The leading term of the ELBO is the Gaussian-convolved objective g. In addition, due to the zero-mean prior, there is a quadratic regularization term on $\boldsymbol{\theta}$, and a term that tries to match the magnitude of λ and $\boldsymbol{\Sigma}$. If we consider the case of an uninformed prior (i.e., limit of large $\boldsymbol{\Sigma}$), then the ELBO simplifies to

$$\text{ELBO}(\boldsymbol{\theta}, \lambda) = -\left[f \star k_\lambda\right](\boldsymbol{\theta}) + \frac{1}{2}m\log(\lambda) - K. \tag{18}$$

Maximizing equation 18 with respect to $\boldsymbol{\theta}$ is then equivalent to minimizing the convolved objective g. Because of the $\log(\lambda)$ term however, it is not guaranteed to be monotonic in λ.

B OPTIMIZING THE CONTINUATION PARAMETER (PROOF OF THEOREM 2)

Before we begin, we establish the following lemma.

Lemma 1. *Let $k_\lambda(\boldsymbol{\theta})$ be a Gaussian kernel with covariance $\boldsymbol{\Sigma} = \lambda \boldsymbol{I}$, as in Section 2.1. Then,*

$$\operatorname{tr}\left(\frac{\partial^2 k_\lambda}{\partial \boldsymbol{\theta}^2}\right) = 2\frac{\partial k_\lambda}{\partial \lambda}. \tag{19}$$

Proof. Starting with the λ derivative,

$$\frac{\partial k_\lambda}{\partial \lambda} = \left(\frac{1}{2\lambda^2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \frac{m}{2\lambda}\right) k_\lambda,$$

Then comparing to the $\boldsymbol{\theta}$ Hessian trace,

$$\begin{aligned}
\frac{\partial^2 k_\lambda}{\partial \boldsymbol{\theta}^2} &= \left(\frac{1}{\lambda^2}\boldsymbol{\theta}\boldsymbol{\theta}^T - \frac{1}{\lambda}\boldsymbol{I}\right) k_\lambda \\
\operatorname{tr}\left(\frac{\partial^2 k_\lambda}{\partial \boldsymbol{\theta}^2}\right) &= \left(\frac{1}{\lambda^2}\operatorname{tr}\left(\boldsymbol{\theta}\boldsymbol{\theta}^T\right) - \frac{1}{\lambda}\operatorname{tr}\left(\boldsymbol{I}\right)\right) k_\lambda \\
&= \left(\frac{1}{\lambda^2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \frac{m}{\lambda}\right) k_\lambda \\
&= 2\frac{\partial k_\lambda}{\partial \lambda}.
\end{aligned}$$

□

To prove monotonicity, we show that $\frac{dg}{d\lambda}(\boldsymbol{\theta}^\star(\lambda), \lambda) > 0$. Expanding out the derivative,

$$\frac{dg}{d\lambda}(\boldsymbol{\theta}^\star(\lambda), \lambda) = \frac{\partial g}{\partial \boldsymbol{\theta}}\frac{\partial \boldsymbol{\theta}^\star}{\partial \lambda}(\boldsymbol{\theta}^\star(\lambda), \lambda) + \frac{\partial g}{\partial \lambda}(\boldsymbol{\theta}^\star(\lambda), \lambda).$$

Because $\boldsymbol{\theta}^\star(\lambda)$ is a stationary point, $\frac{\partial g}{\partial \boldsymbol{\theta}} = 0$, so

$$\frac{dg}{d\lambda}(\boldsymbol{\theta}^\star(\lambda), \lambda) = \frac{\partial g}{\partial \lambda}(\boldsymbol{\theta}^\star(\lambda), \lambda) = \left[f \star \frac{\partial k_\lambda}{\partial \lambda}\right](\boldsymbol{\theta}^\star(\lambda)),$$

and because $\boldsymbol{\theta}^\star(\lambda)$ is a minimum, $\frac{\partial^2 g}{\partial \boldsymbol{\theta}^2}$ is positive semi-definite, so $\operatorname{tr}\left(\frac{\partial^2 g}{\partial \boldsymbol{\theta}^2}\right) > 0$. Expanding the Hessian trace,

$$\begin{aligned}
\operatorname{tr}\left(\frac{\partial^2 g}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta}^\star(\lambda))\right) &= \left[f \star \operatorname{tr}\left(\frac{\partial^2 k_\lambda}{\partial \boldsymbol{\theta}^2}\right)\right](\boldsymbol{\theta}^\star(\lambda)) && \text{by linearity of the convolution} \\
&= 2\left[f \star \frac{\partial k_\lambda}{\partial \lambda}\right](\boldsymbol{\theta}^\star(\lambda)) && \text{by Lemma 1} \\
&= 2\frac{\partial g}{\partial \lambda}(\boldsymbol{\theta}^\star(\lambda), \lambda) > 0.
\end{aligned}$$

This result may be slightly generalized for practical use. Because we showed that for any Gaussian-convolved g,

$$\operatorname{tr}\left(\frac{\partial^2 g}{\partial \boldsymbol{\theta}^2}\right) = 2\frac{\partial g}{\partial \lambda}, \tag{20}$$

g increases monotonically with λ not only when it is at a minimum with respect to $\boldsymbol{\theta}$, but whenever the trace of the Hessian of g with respect to $\boldsymbol{\theta}$ is positive. This implies that as long as we are in a locally convex "valley" of g (for example, by initializing with $(\boldsymbol{\theta}^\star(\lambda_0), \lambda_0)$), then minimizing g with respect to $\boldsymbol{\theta}$ and λ will naturally bring λ toward zero. If $\frac{\partial g}{\partial \lambda}$ ever goes negative in practice, this indicates that g is not locally convex in $\boldsymbol{\theta}$.

C CHANGE OF VARIABLES

Optimizing a loss g with respect to λ would normally require enforcing that λ be nonnegative, however this can be avoided with the substitution $\lambda = \exp(\psi)$. The Monte Carlo gradient estimator for ψ is

$$\begin{aligned}\frac{\partial g}{\partial \psi} = \frac{\partial g}{\partial \lambda}\frac{d\lambda}{d\psi} &\approx \left(\frac{1}{N}\sum_{i=1}^{N}\frac{(\boldsymbol{\theta}-\boldsymbol{\vartheta}_i)^T(\boldsymbol{\theta}-\boldsymbol{\vartheta}_i)-m\lambda}{2\lambda^2}f(\boldsymbol{\vartheta}_i)\right)(\exp(\psi)) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{(\boldsymbol{\theta}-\boldsymbol{\vartheta}_i)^T(\boldsymbol{\theta}-\boldsymbol{\vartheta}_i)-m\exp(\psi)}{2\exp(\psi)}f(\boldsymbol{\vartheta}_i),\end{aligned} \tag{21}$$

where the estimator for $\frac{\partial g}{\partial \lambda}$ is from equation 8. We can also reparameterize the random samples by

$$\boldsymbol{\theta}-\boldsymbol{\vartheta}_i = \sqrt{\lambda}\boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_N \sim \mathcal{N}(\mathbf{0},\boldsymbol{I}), \tag{22}$$

which allows us to cancel some terms, leading to

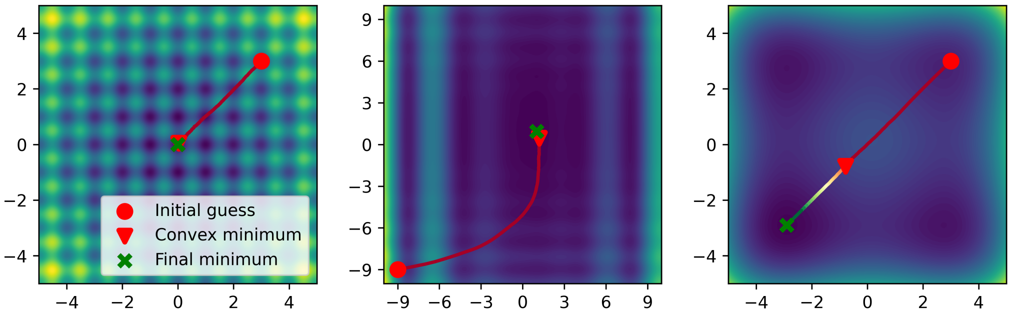
$$\frac{\partial g}{\partial \psi} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i - m}{2}f\left(\boldsymbol{\theta}-\sqrt{\exp(\psi)}\boldsymbol{\epsilon}_i\right). \tag{23}$$

If we include the regularization term from Section 2.5,

$$\begin{aligned}\frac{\partial g}{\partial \psi} = \frac{\partial g}{\partial \lambda}\frac{d\lambda}{d\psi} &\approx \left(\frac{1}{N}\sum_{i=1}^{N}\frac{(\boldsymbol{\theta}-\boldsymbol{\vartheta}_i)^T(\boldsymbol{\theta}-\boldsymbol{\vartheta}_i)-m\lambda}{2\lambda^2}f(\boldsymbol{\vartheta}_i)+\beta m\right)(\exp(\psi)) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i - m}{2}f\left(\boldsymbol{\theta}-\sqrt{\exp(\psi)}\boldsymbol{\epsilon}_i\right)+\beta m\exp(\psi).\end{aligned} \tag{24}$$

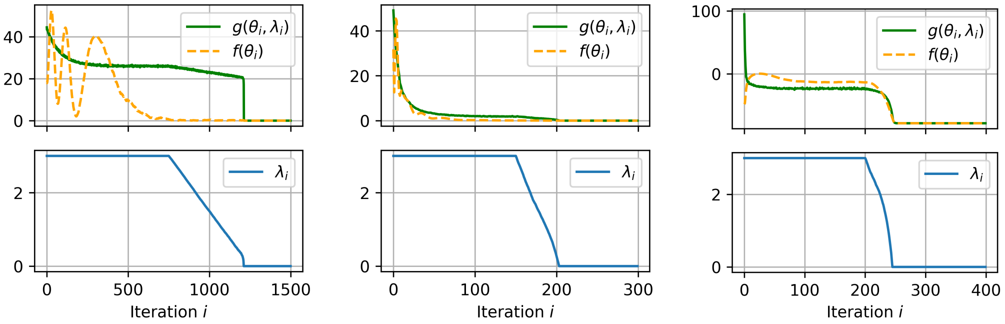
To decouple the scaling of β from λ, we scale the regularization term by the initial λ_0 (i.e., $\exp(\psi_0)$) without loss in generality, giving

$$\frac{\partial g}{\partial \psi} = \frac{1}{N}\sum_{i=1}^{N}\frac{\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i - m}{2}f\left(\boldsymbol{\theta}-\sqrt{\exp(\psi)}\boldsymbol{\epsilon}_i\right)+\beta m\exp(\psi-\psi_0). \tag{25}$$



(a) Rastrigin (b) Levy (c) Styblinski-Tang

Figure 8: Optimization path on all 2D test functions. The colour of the trace corresponds to the value of λ.



(a) Rastrigin (b) Levy (c) Styblinski-Tang

Figure 9: Optimization by continuation on all 2D test functions with large number of Monte Carlo samples.

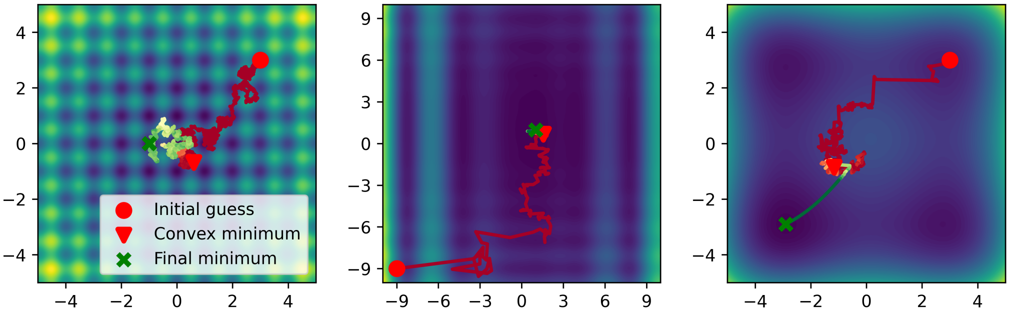
D 2D TEST FUNCTIONS

The non-convex functions used in this section, Rastrigin, Levy, and Styblinski-Tang, were chosen for having many local minima and saddle points. For ease of visualization, all example test functions are 2D. In all test cases, we use simple gradient descent with $\lambda_0 = 3$. The optimizer is run for a "warmup period," meaning λ is held constant at λ_0 for a certain number of iterations, after which λ is also adapted by gradient descent. The learning rates for $\boldsymbol{\theta}$ and λ, the warmup period, and the total number of iterations are varied for each test case. All case-specific information is given in Appendix E.

D.1 MANY MONTE CARLO SAMPLES

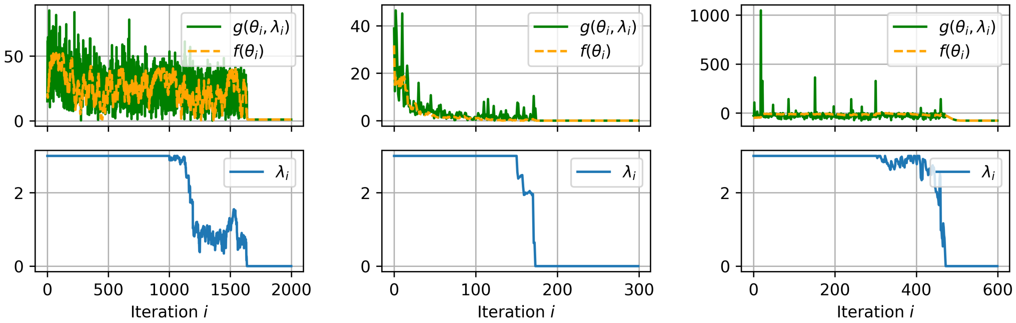
This scenario is an "ideal" case, where large numbers of Monte Carlo samples are used to estimate g and ∇g by equation 7 and equation 8. This gives relatively low-variance estimates of these quantities at each optimizer step, and this is reflected in the smooth descent towards each function's respective minimum.

Surface plots of each test function are given in Figure 8. They also show the optimization path from each initial guess, first to the "convex minimum" (i.e. the minimum of g at λ_0), which corresponds to the warmup period, then to the final minimum as λ decreases to zero. The Rastrigin and Levy functions behave similarly in that most of the optimization occurs during the warmup period, as their respective convex minima closely coincide with their true minima. The Styblinski-Tang function however shows how continuation behaves when the local minima are more even. The convex minimum is close to the midpoint between the four local minima. As λ decreases, the optimizer approaches the true minimum in the bottom left corner.



(a) Rastrigin (b) Levy (c) Styblinski-Tang

Figure 10: Optimization path on all 2D test functions. The colour of the trace corresponds to the value of λ.



(a) Rastrigin (b) Levy (c) Styblinski-Tang

Figure 11: Optimization by continuation on all 2D test functions with one Monte Carlo sample.

The convergence plots of f, g, and λ over iteration number are shown in Figure 9. The values along these plots correspond to the optimization path from Figure 8. In all the convergence plots, the fact that continuation allows the optimizer to climb out of local minima is apparent in the non-monotonic behaviour of f compared to the monotonic behaviour of g. The convex minimum is reached as g plateaus, at which point the warmup period concludes and λ is allowed to decrease. When $\lambda = 0$, $g = f$.

D.2 ONE MONTE CARLO SAMPLE

This scenario is a more realistic case. To minimize the difference in computational cost between continuation and standard gradient-based optimization, only one Monte Carlo sample is used to estimate g and ∇g. In this special case, continuation amounts to adding Gaussian noise of diminishing variance to $\boldsymbol{\theta}$ during optimization.

Surface plots of each test function with optimization path are given in Figure 10. The effect of the additional noise is apparent; all but Levy require a reduced learning rate and increased number of iterations, and the Rastrigin optimizer settles in the wrong local minimum regardless. In the ideal scenario shown previously, the warmup period eliminates variability in the final minimum due to the initial guess for $\boldsymbol{\theta}$, since $\boldsymbol{\theta}$ was optimized to a unique convex minimum. In this scenario it still serves the same role, however the convex minimum is now a random variable drawn from a potentially high-variance distribution. This is best exemplified with Rastrigin, as it resulted in the optimizer finding an adjacent local minimum.

The convergence plots of f, g, and λ over iteration number are shown in Figure 11. The noisy optimization is evident in all the convergence plots, especially in the slower descent of λ. Once $\lambda = 0$ however, the optimizer is performing simple gradient descent.

E 2D TEST FUNCTIONS AND OPTIMIZATION SETUPS

This section summarizes the test functions and their respective optimizer setups shown in Appendix D and the saddle function in Section 2.4. The test functions are the Rastrigin function,

$$f(\boldsymbol{\theta}) = 10m + \sum_{i=1}^{m} \left(\theta_i^2 - 10\cos(2\pi\theta_i)\right), \tag{26}$$

the Levy function,

$$\begin{aligned} f(\boldsymbol{\theta}) &= \sin^2(\pi w_1) + (w_m - 1)^2 \left(1 + \sin^2(2\pi w_m)\right) \\ &+ \sum_{i=1}^{m-1} (w_i - 1)^2 \left(1 + 10\sin^2(\pi w_i + 1)\right), \\ &\text{where} \quad w_i = 1 + \frac{\theta_i - 1}{4} \quad \text{for all } i = 1, \ldots, m, \end{aligned} \tag{27}$$

and the Styblinski-Tang function,

$$f(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{m}(\theta_i^4 - 16\theta_i^2 + 5\theta_i). \tag{28}$$

The saddle function is defined in Section 2.4. For all test functions, the parameter dimension m is 2. The true minima for each are listed in Table 1.

Table 1: Non-convex 2D test functions

Function	Minimizer $\boldsymbol{\theta}^\star$	Minimum $f(\boldsymbol{\theta}^\star)$
Saddle, $a = 0.01$	$\pm(1.9222, 1.9222)$	-0.3431
Saddle, $a = 0.02$	$\pm(1.8451, 1.8451)$	-0.1920
Rastrigin	$(0, 0)$	0
Levy	$(1, 1)$	0
Styblinski-Tang	$(-2.9035, -2.9035)$	-78.3320

For each case, simple gradient descent was used with separate learning rates for $\boldsymbol{\theta}$ and λ. The change of variables (Appendix C) and regularization (Section 2.5) were not used for these test cases; they were introduced only for the scenario of large machine learning problems.

The hyperparameter values for each case and figure in Section 2.4 and Appendix D are summarized in Table 2.

Table 2: Optimization hyperparameters for non-convex 2D test functions

Figure	Function	Monte Carlo samples	Learning rate on $\boldsymbol{\theta}$	Learning rate on λ	Warmup / Total # iter.
4b	Saddle, $a = 0.01$	4096	3×10^{0}	3×10^{0}	200/600
4a	Saddle, $a = 0.02$	4096	3×10^{0}	3×10^{0}	200/600
8a, 9a	Rastrigin	4096	3×10^{-3}	3×10^{-3}	750/1500
8b, 9b	Levy	1024	1×10^{-1}	1×10^{-1}	150/300
8c, 9c	Styblinski-Tang	4096	1×10^{-2}	1×10^{-2}	200/400
10a, 11a	Rastrigin	1	1×10^{-3}	3×10^{-3}	1000/2000
10b, 11b	Levy	1	1×10^{-1}	1×10^{-1}	150/300
10c, 11c	Styblinski-Tang	1	3×10^{-3}	1×10^{-2}	300/600

F DYNAMICAL SYSTEMS

The first learning problem is the Lotka-Volterra system of predator and prey populations,

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha x - \beta xy \\ \delta xy - \gamma y \end{bmatrix}, \tag{29}$$

which is parameterized by α, β, γ, and δ. To form the dataset, β and δ are held constant at $4/3$ and 1 respectively, and α and γ are varied according to $\alpha \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ and $\gamma \in \{1, 1.25, 1.5, 1.75, 2\}$, leading to 30 different parameter instances and 30 corresponding time series. The initial condition is fixed at $[x_0, y_0] = [1, 1]$. Each time series consists of solutions to equation 29 at 50 evenly spaced time slices in $t \in [0, 20]$.

The second learning problem is the Lorenz system,

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \sigma(y - x) \\ x(\rho - z) - y \\ xy - \beta z \end{bmatrix}, \tag{30}$$

which is parameterized by σ, ρ, and β. For this dataset, β is held constant at $8/3$, and σ and ρ are varied according to $\sigma \in \{9, 9.5, 10, 10.5, 11\}$ and $\rho \in \{25, 26, 27, 28, 29, 30\}$, leading again to 30 different parameter instances and 30 corresponding time series. The initial condition is fixed at $[x_0, y_0, z_0] = [5, 5, 5]$. Each time series consists of solutions to equation 30 at 10 evenly spaced time slices in $t \in [0, 1]$.