Inferring Black Hole Properties from Astronomical Multivariate Time Series with Bayesian Attentive Neural Processes

Ji Won Park [1 2] **Ashley Villar** [3 4] **Yin Li** [4] **Yan-Fei Jiang** [4] **Shirley Ho** [4 5 6 7] **Joshua Yao-Yu Lin** [8]
Phil Marshall [1 2] **Aaron Roodman** [1 2]

Abstract

Among the most extreme objects in the Universe, active galactic nuclei (AGN) are luminous centers of galaxies where a black hole feeds on surrounding matter. The variability patterns of the light emitted by an AGN contain information about the physical properties of the underlying black hole. Upcoming telescopes will observe over 100 million AGN in multiple broadband wavelengths, yielding a large sample of multivariate time series with long gaps and irregular sampling. We present a method that reconstructs the AGN time series and simultaneously infers the posterior probability density distribution (PDF) over the physical quantities of the black hole, including its mass and luminosity. We apply this method to a simulated dataset of 11,000 AGN and report precision and accuracy of 0.4 dex and 0.3 dex in the inferred black hole mass. This work is the first to address probabilistic time series reconstruction and parameter inference for AGN in an end-to-end fashion.

1. Introduction

Supermassive black holes (BHs) reside at the centers of most galaxies, feeding on diffuse matter around them. Radiation from matter falling into their gravitational pull makes these galactic centers, called active galactic nuclei (AGN), some of the most luminous in the Universe. The time variability patterns of AGN light are correlated with the physical properties of the underlying BH, such as the mass, rate of matter inflow, and age (Wold et al., 2008; Simm et al., 2016; MacLeod et al., 2010; Suberlak et al., 2021).

Being so luminous, AGN can be observed out to great distances, close to the edge of the observable Universe (e.g., Mortlock et al. 2011). Characterizing faraway BHs gives us a glimpse into the little-known early Universe. By inferring BH physics from AGN light, we can gain an understanding the origin and evolution of the cosmos, including the nature of dark energy and dark matter (Khadka & Ratra, 2020).

Upcoming large-sky telescope surveys herald an unprecedented increase in the AGN data volume. The Vera Rubin Observatory Legacy Survey of Space and Time (LSST) is projected to yield 100 million AGN time series in six optical broadband filters over ten years (Abell et al., 2009). Traditional methods of estimating BH properties, however, rely on expensive spectroscopic data, i.e. measurements of the AGN light at continuous wavelengths. Obtaining spectroscopy for millions of objects would be unfeasible. To take advantage of all the new data, we require an efficient method that can estimate desired quantities directly from the 6-filter LSST time series. The learned relationship will improve our understanding of BH physics, particularly as a physical model of AGN variability does not exist.

We present a method that simultaneously reconstructs the multivariate AGN time series from limited observations and infers the full posterior probability density distribution (PDF) over key BH properties. Our method is designed for irregular, multivariate time series, as telescope data often suffer from long seasonal gaps and irregular sampling as well as noise from the Earth's atmosphere and telescope optics. Uncertainty quantification is essential to optimize follow up strategies.

This work is the first deep learning pipeline that simultaneously addresses AGN light curve reconstruction and parameter inference in a probabilistic manner. It is additionally the first designed for multivariate time series. At the core of our pipeline is an attentive neural process (Kim et al., 2019), a type of latent variable model, that has been modified for density estimation. In the past, autoencoders have been applied to output point estimates of the unobserved portions

[1]Kavli Institute for Particle Astrophysics and Cosmology, Department of Physics, Stanford University, Stanford, CA, USA [2]SLAC National Accelerator Laboratory, Menlo Park, CA, USA [3]Department of Astronomy, Columbia University, New York, NY, USA [4]Flatiron Institute, New York, NY, USA [5]Princeton University, Princeton,NJ 08540 [6]New York university, New York, NY 10010 [7]Carnegie Mellon University, Pittsburgh, PA 15289 [8]University of Illinois at Urbana-Champaign, Champaign, IL, USA. Correspondence to: Ji Won Park <jwp@stanford.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

of the light curve in a single-filter setting (Tachibana et al., 2020). Convolutional neural nets were trained to point-estimate the redshifts based on multi-filter light curves from the Sloan Digital Sky Survey (SDSS) (Schneider et al., 2010; Pasquet-Itam & Pasquet, 2018). Summary statistics from the SDSS light curves were also fed into a neural net for point-estimating BH mass and redshift (Lin et al., 2020).

2. Data

2.1. Multi-filter time series

The training set consisted of 11,000 simulated multi-filter time series and the corresponding target labels. The validation set contained 50 held-out examples, drawn from the same distribution as the training set. The input was a simulated time series of the AGN flux, or brightness, in the six bandpass filters $ugrizY$. This six-dimensional light curve followed the Ornstein-Uhlenbeck (OU) process, a widely adopted stochastic model of AGN flux variability (Kelly et al., 2009). The light curves were sampled at irregular times, with long gaps, to simulate LSST-like observations (Reuter et al., 2016). We added astrophysical noise to the light curves on the fly during training (Kessler et al., 2019).

2.2. Target quantities

The target quantities were taken from the Second Data Challenge (DC2), a simulated LSST-like catalog (LSST Dark Energy Science Collaboration et al., 2021a;b). For each of the $ugrizY$ filters, there were three variability parameters: the average flux (m), long-term amplitude (SF_∞) of fluctuations, and characteristic timescale (τ) of fluctuations. The SF_∞ and τ parameters can be interpreted as the maximum amplitude of flux fluctuations and the timescale to reach such an amplitude, respectively. We also included the BH mass, redshift (a measure of distance), and the i-band absolute magnitude (M_i; a measure of the intrinsic luminosity of the AGN). The 21 quantities in total shared correlations empirically modeled after a well-known dataset (SDSS) of quasars (Kelly et al., 2009; MacLeod et al., 2010).

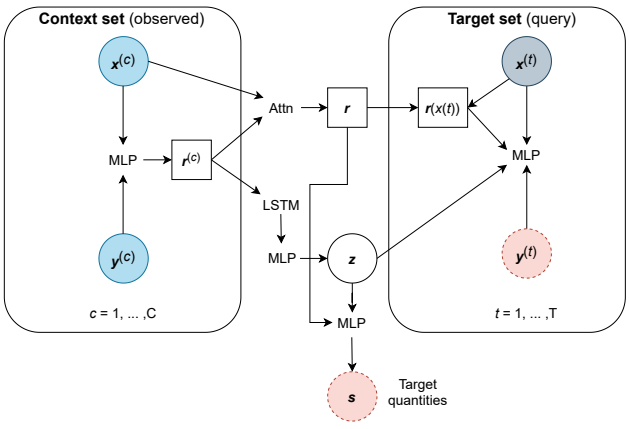

Figure 1. Model architecture of the Bayesian ANP.

3. Method

3.1. Model

See Figure 1 for a diagram of our model architecture. We used a latent attentive neural process (ANP) (Kim et al., 2019) to model the conditional distribution over regression functions mapping our input times, $\boldsymbol{x}^{(c)} \in \mathbb{R}$, to our output $ugrizY$ fluxes, $\boldsymbol{y}^{(c)} \in \mathbb{R}^6$. Conditioning on observed time/flux pairs, $(\boldsymbol{x}_C, \boldsymbol{y}_C) := \left(\boldsymbol{x}^{(c)}, \boldsymbol{y}^{(c)}\right)_{c \in C}$, called the *context* set, we query the model for the fluxes at some unobserved times, $(\boldsymbol{x}_T, \boldsymbol{y}_T) := \left(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(c)}\right)_{c \in C}$, called the *target* set. A common convention, used here, is to set $C \subset T$.

In the *deterministic* path of the forward model, each time/flux pair $\boldsymbol{x}^{(c)}, \boldsymbol{y}^{(c)}$ passes through an MLP to form a representation $\boldsymbol{r}^{(c)} \in \mathbb{R}^h$. Self-attention is applied to the resulting context representations and the model attends to these via cross-attention with $\boldsymbol{x}^{(t)}$ to predict $\boldsymbol{y}^{(t)}$. The attention mechanism allows it to attribute higher relative importance to time samples that are more informative for prediction. This is useful for astronomical data, where observations tend to be clustered together and some carry higher signal-to-noise than others.

The latent path encodes a global understanding of the entire time series. The encodings $\boldsymbol{r}_C := \left(\boldsymbol{r}^{(c)}\right)_{c \in C}$ are aggregated using an LSTM and passed through a multi-layer perceptron (MLP) to form a latent variable, $\boldsymbol{z}$.[1] We model $\boldsymbol{z}$ as a factorized Gaussian, with a Gaussian prior. Each sample of $\boldsymbol{z}$ represents one realization of the data-generating stochastic process, so $\boldsymbol{z}$ stores information about uncertainty in $\boldsymbol{y}_T$.

As an update to the original ANP architecture, we additionally use $\boldsymbol{z}$ and the mean of the attention-reweighted vector $\boldsymbol{r}$ to infer the global target quantities $\boldsymbol{s} \in \mathbb{R}^{21}$, described in Section 2.2. An MLP takes in $\boldsymbol{z}$ and outputs the parameters defining the posterior PDF over $\boldsymbol{s}$.

3.2. Uncertainty quantification

We adapted the standard ANP into a Bayesian ANP (BANP) to enable posterior inference over the network weights (Denker & LeCun, 1991). Our posterior on $\boldsymbol{y}_T, \boldsymbol{s}$ was:

$$p(\boldsymbol{y}_T, \boldsymbol{s} | \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C) = \int p(\boldsymbol{y}_T, \boldsymbol{s} | \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, W) p(W | \boldsymbol{x}_C, \boldsymbol{y}_C)\, dW, \quad (1)$$

where W denotes the network weights. The likelihood $p(\boldsymbol{y}_T, \boldsymbol{s} | \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, W)$ captures the *aleatoric* uncertainty, which exists due to the intrinsic randomness in the data-generating process, e.g. the spread in the correlation between luminosity and BH mass. We marginalize over the weight posterior $p(W | \boldsymbol{x}_C, \boldsymbol{y}_C)$ to account for the *epistemic* uncertainty, which originates from incomplete knowledge,

[1] The ANP, as originally proposed by Kim et al. 2019, uses mean aggregation but we find that the LSTM yield better predictions of our target quantities.

e.g. limited training data.

Aleatoric uncertainty: For predicting $\boldsymbol{y}_T$, we would maximize the latent neural process likelihood (Garnelo et al., 2018):

$$p\left(\boldsymbol{y}_T|\boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, W\right) = \int p\left(\boldsymbol{y}_T|\boldsymbol{x}_T, \boldsymbol{r}_C, \boldsymbol{z}, W\right) q\left(\boldsymbol{z}|\boldsymbol{x}_C, \boldsymbol{y}_C, W\right) d\boldsymbol{z}, \quad (2)$$

where we have explicitly conditioned on W. Similarly, the likelihood of $\boldsymbol{s}$ is

$$p\left(\boldsymbol{s}|\boldsymbol{x}_C, \boldsymbol{y}_C, W\right) = \int p\left(\boldsymbol{s}|\boldsymbol{z}, W\right) q\left(\boldsymbol{z}|\boldsymbol{x}_C, \boldsymbol{y}_C, W\right) d\boldsymbol{z}. \quad (3)$$

We chose $p(\boldsymbol{s}|\boldsymbol{z})$ to be multivariate Gaussian with a full covariance matrix, so as to model physical correlations between the target quantities. The total aleatoric portion of our likelihood combines Equations 2 and 3:

$$\log p(\boldsymbol{y}_T, \boldsymbol{s}|\boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, W) \propto \alpha \log p\left(\boldsymbol{y}_T|\boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, W\right) + \log p\left(\boldsymbol{s}|\boldsymbol{x}_C, \boldsymbol{y}_C, W\right), \quad (4)$$

where α is a hyperparameter controlling the relative weight between light curve reconstruction and parameter inference.

Epistemic uncertainty: We used Monte Carlo (MC) dropout (Gal & Ghahramani, 2016; Kendall & Gal, 2017) to compute the integral in Equation 1. MC dropout replaces the true weight posterior with a simple Bernoulli variational approximation $q(W|\boldsymbol{x}_C, \boldsymbol{y}_C)$, which can be implemented by setting random network weights to zero during training and testing. We treated the dropout rate as a hyperparameter.

3.3. Optimization

Context-target split: We provided as context the observed portions of the light curve according to the 10-year LSST observing strategy, which consisted of 500-1,000 samplings per filter. All six filters were observed at the same times. The target was the union of the context set and the light curve at every 10 days, i.e. 365 samplings per filter.

Loss function: The exact form of Equation 4 is intractable. The evidence lower bound (ELBO) can be optimized, however, using the reparameterization trick (Kingma & Welling, 2013; Rezende et al., 2014):

$$\log p(\boldsymbol{y}_T, \boldsymbol{s}|\boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, W) \geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_T, \boldsymbol{y}_T)}\left[\alpha' \log p(\boldsymbol{y}_T|\boldsymbol{x}_T, \boldsymbol{r}_C, \boldsymbol{z}, W) + \log p(\boldsymbol{s}|\boldsymbol{z}, W)\right] - D_{\mathrm{KL}}\left(q(\boldsymbol{z}|\boldsymbol{x}_T, \boldsymbol{y}_T, W)||q(\boldsymbol{z}|\boldsymbol{x}_C, \boldsymbol{y}_C, W)\right) := \mathcal{F}_{\mathrm{ELBO}}(W) \quad (5)$$

where α' serves the same reweighting purpose as α in Equation 4. We had $\alpha' = 10$. In each training iteration, given a realization of weights from MC dropout, $\hat{W} \sim q(\hat{W}|\boldsymbol{x}_C, \boldsymbol{y}_C)$, we performed gradient descent on $\mathcal{F}_{\mathrm{ELBO}}(\hat{W})$ with a weight decay of 1e-5 using the ADAM optimizer (Kingma & Ba, 2014). Training was done for 300 epochs in batch sizes of 40. The learning rate began with 1e-3 and was halved whenever the validation loss did not decrease for 20 epochs.

Table 1. Validation metrics of time series reconstruction and parameter recovery. Lower is better. Error bars indicate the standard deviations over three training random seeds.

MODEL	RECONSTRUCTION		PARAMETER	
	1-σ	MAE	1-σ	MAE
BANP	3.1 ± 0.9	2.2 ± 0.8	$\mathbf{51 \pm 2}$	$\mathbf{46 \pm 1}$
BASELINE	N/A	N/A	89 ± 3	97 ± 3

4. Results

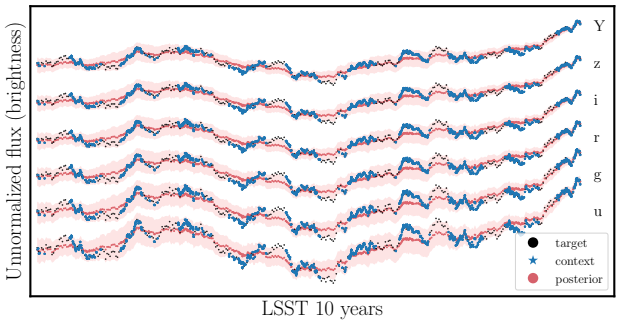

Figure 2. Reconstructed $ugrizY$ light curve for a validation AGN. Blue stars are observed *context* points, provided to the network at test time. Black dots are query *target* points, for which the network produces the red posteriors. The true observations lie within the 1-σ region of our predicted light curves.

4.1. Light curve reconstruction

Figure 2 shows the BANP posterior for a validation $ugrizY$ light curve. Our model closely reconstructs the long-term amplitude and timescale of fluctuations. This performance indicates a good understanding of SF_∞ and τ. While the truth is consistent with its 1-σ credible interval at all times, the uncertainties are overestimated and the model does a poorer job of predicting the shorter-term fluctuations. One possibility is that the reconstruction loss competes with the parameter inference loss, as the model does not have to get the shorter-term fluctuations correct to predict SF_∞ and τ.

To assess the precision of reconstruction, we draw 100 light curves from our flux posterior and approximate the 1-σ credible width at each target time as the standard deviation across the samples. We then take the average across all target times and 50 validation AGN. Similarly, to assess the accuracy, we obtain the absolute error of the central flux prediction at each target time and take the average across all target times in the validation set (mean absolute error; MAE). The resulting values, which carry units of magnitude, are listed in Table 1.

4.2. Retrieval of target quantities

Figure 3 shows the inferred posterior over a representative subset of the target quantities in $\boldsymbol{s}$, for a single AGN. Our

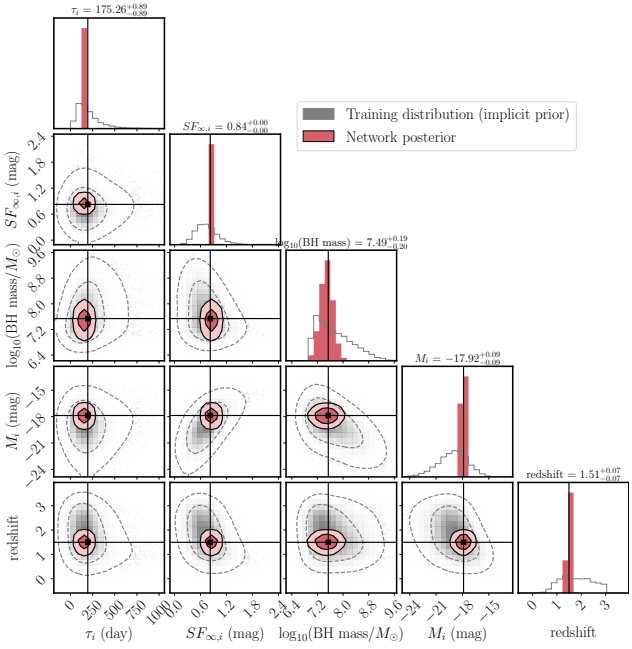

Figure 3. Network posterior for a validation AGN overlaid with the training distribution for a representative subset of the target quantities. The vertical/horizontal black lines show the true parameter values. In this example, our network produces a tight and unbiased estimate for the BH and light curve parameters.

BANP can accurately recover the true quantities within 1-σ credible interval.

To assess the precision and accuracy of our parameter recovery, we proceed similarly as in Section 4.1 to evaluate the 1-σ uncertainty and MAE of our posterior on the 21 target quantities. The resulting values are listed on Table 1. Our baseline method was a residual MLP that took as input the summary statistics of the light curve—the mean and standard deviation of the flux in six filters—to yield an identically parameterized posterior. We find a twofold improvement in target recovery, in both precision and accuracy. In particular, the timescale-sensitive quantities τ and BH mass showed the most improvement, of 150% and 120%, respectively. The BH mass was precise to 0.4 dex.

4.2.1. Calibration

To assess the calibration of our parameter inference, we plot the metric introduced in Wagner-Carena et al. (2021) in Figure 4.[2] For a given fraction of the BANP posterior probability volume, p_X, the metric plots the fraction of posterior samples that contain the truth within the volume, p_Y. If the posterior is perfectly calibrated, p_X of the posterior samples would encompass the truth $p_Y = p_X$ of the time, for every value of p_X. We apply this metric on the validation set as a whole by averaging the p_Y values across the validation AGN, to get $p_Y^{\rm val}$. Regions of the curve with $p_Y^{\rm val} < p_X$ indicate overconfidence whereas the opposite indicates underconfidence. We find good calibration for our choice of MC dropout rate (0.05).

[2]The supplementary material describes this metric in detail.

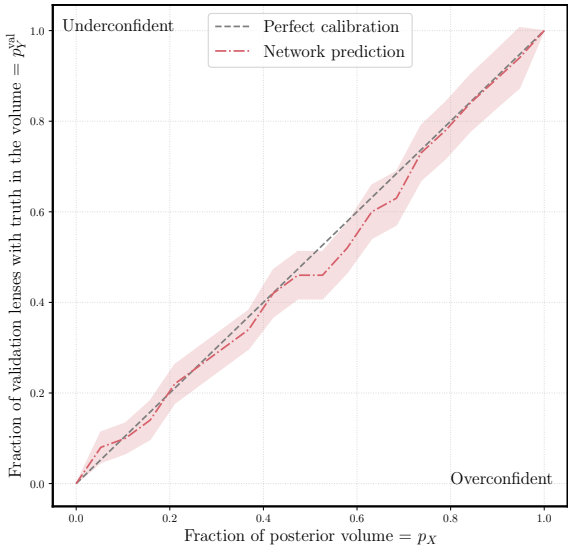

Figure 4. The confidence of our network's parameter inference is statistically consistent with the truth. The calibration metric for our model falls on the $p_Y^{\rm val} = p_X$ line, which indicates perfect calibration. Error bands are estimated using jacknife sampling of the posterior samples.

5. Discussion and Future Work

We have adapted latent neural processes for probabilistic parameter estimation and multivariate time series reconstruction. Our method is capable of interpolating AGN light curves and precisely inferring BH parameters. Compared to classical fitting methods, the BANP is flexible and does not assume a fixed parameterization or kernel. This is important for processing real AGN data, because there exists no physical parameterized model that can reliably describe the temporal patterns. Our method is thus useful for extracting an informative latent space from noisy and irregularly sampled time series in general.

In future work, we plan to upgrade to more physical simulations and asynchronous sampling in $ugrizY$, in preparation for real LSST AGN data. Ultimately, we aspire to hierarchically infer the hyperparameters that govern the AGN population, using the constraints from the individual AGN. Likelihood-free inference methods such as normalizing flows (Rezende & Mohamed, 2015) are interesting for a flexible posterior parameterization. We will also explore incorporating physics priors into the network architecture; scalable variants of Gaussian processes (e.g. Salimbeni et al. 2017, Wilson et al. 2016) and latent stochastic differential equations (Li et al., 2020) are promising alternatives to the ANP for modeling temporal correlations.

Acknowledgements

We thank Xuechen Li and David Duvenaud for the insightful discussions on joint reconstruction and parameter inference using latent variable models.

References

Abell, P. A., Burke, D. L., Hamuy, M., Nordby, M., Axelrod, T. S., Monet, D., Vrsnak, B., Thorman, P., Ballantyne, D., Simon, J. D., et al. Lsst science book, version 2.0. Technical report, 2009.

Denker, J. S. and LeCun, Y. Transforming neural-net output levels to probability distributions. In *NeurIPS*, pp. 853–859, 1991.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pp. 1050–1059, 2016.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

Kelly, B. C., Bechtold, J., and Siemiginowska, A. Are the variations in quasar optical flux driven by thermal fluctuations? *The Astrophysical Journal*, 698(1):895, 2009.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, pp. 5574–5584, 2017.

Kessler, R., Narayan, G., Avelino, A., Bachelet, E., Biswas, R., Brown, P., Chernoff, D., Connolly, A., Dai, M., Daniel, S., et al. Models and simulations for the photometric lsst astronomical time series classification challenge (plasticc). *Publications of the Astronomical Society of the Pacific*, 131(1003):094501, 2019.

Khadka, N. and Ratra, B. Using quasar x-ray and uv flux measurements to constrain cosmological model parameters. *Monthly Notices of the Royal Astronomical Society*, 497(1):263–278, 2020.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Li, X., Wong, T.-K. L., Chen, R. T., and Duvenaud, D. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

Lin, J. Y.-Y., Pandya, S., Pratap, D., Liu, X., and Kind, M. C. Agnet: Weighing black holes with machine learning. *arXiv preprint arXiv:2011.15095*, 2020.

LSST Dark Energy Science Collaboration, Abolfathi, B., Alonso, D., Armstrong, R., Aubourg, É., Awan, H., Babuji, Y. N., Bauer, F. E., Bean, R., Beckett, G., Biswas, R., Bogart, J. R., Boutigny, D., Chard, K., Chiang, J., Claver, C. F., Cohen-Tanugi, J., Combet, C., Connolly, A. J., Daniel, S. F., Digel, S. W., Drlica-Wagner, A., Dubois, R., Gangler, E., Gawiser, E., Glanzman, T., Gris, P., Habib, S., Hearin, A. P., Heitmann, K., Hernandez, F., Hložek, R., Hollowed, J., Ishak, M., Ivezić, Ž., Jarvis, M., Jha, S. W., Kahn, S. M., Kalmbach, J. B., Kelly, H. M., Kovacs, E., Korytov, D., Krughoff, K. S., Lage, C. S., Lanusse, F., Larsen, P., Guillou, L. L., Li, N., Longley, E. P., Lupton, R. H., Mandelbaum, R., Mao, Y.-Y., Marshall, P., Meyers, J. E., Moniez, M., Morrison, C. B., Nomerotski, A., O'Connor, P., Park, H., Park, J. W., Peloton, J., Perrefort, D., Perry, J., Plaszczynski, S., Pope, A., Rasmussen, A., Reil, K., Roodman, A. J., Rykoff, E. S., Sánchez, F. J., Schmidt, S. J., Scolnic, D., Stubbs, C. W., Tyson, J. A., Uram, T. D., Villarreal, A., Walter, C. W., Wiesner, M. P., Wood-Vasey, W. M., and Zuntz, J. The LSST DESC DC2 simulated sky survey. *The Astrophysical Journal Supplement Series*, 253(1):31, mar 2021a. doi: 10.3847/1538-4365/abd62c. URL https://doi.org/10.3847/1538-4365/abd62c.

LSST Dark Energy Science Collaboration, Abolfathi, B., Armstrong, R., Awan, H., Babuji, Y. N., Bauer, F. E., Beckett, G., Biswas, R., Bogart, J. R., Boutigny, D., Chard, K., Chiang, J., Cohen-Tanugi, J., Connolly, A. J., Daniel, S. F., Digel, S. W., Drlica-Wagner, A., Dubois, R., Gawiser, E., Glanzman, T., Habib, S., Hearin, A. P., Heitmann, K., Hernandez, F., Hložek, R., Hollowed, J., Jarvis, M., Jha, S. W., Kalmbach, J. B., Kelly, H. M., Kovacs, E., Korytov, D., Krughoff, K. S., Lage, C. S., Lanusse, F., Larsen, P., Li, N., Longley, E. P., Lupton, R. H., Mandelbaum, R., Mao, Y.-Y., Marshall, P., Meyers, J. E., Park, J. W., Peloton, J., Perrefort, D., Perry, J., Plaszczynski, S., Pope, A., Rykoff, E. S., Sánchez, F. J., Schmidt, S. J., Uram, T. D., Villarreal, A., Walter, C. W., Wiesner, M. P., and Wood-Vasey, W. M. Desc dc2 data release note. *arXiv preprint arXiv:2101.04855*, 2021b.

MacLeod, C. L., Ivezić, Ž., Kochanek, C., Kozłowski, S., Kelly, B., Bullock, E., Kimball, A., Sesar, B., Westman, D., Brooks, K., et al. Modeling the time variability of sdss stripe 82 quasars as a damped random walk. *The Astrophysical Journal*, 721(2):1014, 2010.

Mortlock, D. J., Warren, S. J., Venemans, B. P., Patel, M., Hewett, P. C., McMahon, R. G., Simpson, C., Theuns, T., Gonzáles-Solares, E. A., Adamson, A., et al. A luminous

quasar at a redshift of z= 7.085. *Nature*, 474(7353):616–619, 2011.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Park, J. W., Wagner-Carena, S., Birrer, S., Marshall, P. J., Lin, J. Y.-Y., Roodman, A., LSST Dark Energy Science Collaboration, et al. Large-scale gravitational lens modeling with bayesian neural networks for accurate and precise inference of the hubble constant. *The Astrophysical Journal*, 910(1):39, 2021.

Pasquet-Itam, J. and Pasquet, J. Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the sloan digital sky survey stripe 82. *Astronomy & Astrophysics*, 611:A97, 2018.

Reuter, M. A., Cook, K. H., Delgado, F., Petry, C. E., and Ridgway, S. T. Simulating the LSST OCS for conducting survey simulations using the LSST scheduler. 9911:794 – 801, 2016. doi: 10.1117/12.2232680. URL https://doi.org/10.1117/12.2232680.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538. PMLR, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Salimbeni, H. and Deisenroth, M. Doubly stochastic variational inference for deep gaussian processes. *arXiv preprint arXiv:1705.08933*, 2017.

Schneider, D. P., Richards, G. T., Hall, P. B., Strauss, M. A., Anderson, S. F., Boroson, T. A., Ross, N. P., Shen, Y., Brandt, W., Fan, X., et al. The sloan digital sky survey quasar catalog. v. seventh data release. *The Astronomical Journal*, 139(6):2360, 2010.

Simm, T., Salvato, M., Saglia, R., Ponti, G., Lanzuisi, G., Trakhtenbrot, B., Nandra, K., and Bender, R. Pan-starrs1 variability of xmm-cosmos agn-ii. physical correlations and power spectrum analysis. *Astronomy & Astrophysics*, 585:A129, 2016.

Suberlak, K. L., Ivezić, Ž., and MacLeod, C. Improving damped random walk parameters for sdss stripe 82 quasars with pan-starrs1. *The Astrophysical Journal*, 907 (2):96, 2021.

Tachibana, Y., Graham, M. J., Kawai, N., Djorgovski, S., Drake, A. J., Mahabal, A. A., and Stern, D. Deep modeling of quasar variability. *The Astrophysical Journal*, 903 (1):54, 2020.

Wagner-Carena, S., Park, J. W., Birrer, S., Marshall, P. J., Roodman, A., Wechsler, R. H., Collaboration, L. D. E. S., et al. Hierarchical inference with bayesian neural networks: An application to strong gravitational lensing. *The Astrophysical Journal*, 909(2):187, 2021.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Stochastic variational deep kernel learning. *arXiv preprint arXiv:1611.00336*, 2016.

Wold, M., Brotherton, M., and Shang, Z. Black hole mass and variability in quasars. In *AIP Conference Proceedings*, volume 1053, pp. 55–58. American Institute of Physics, 2008.

Supplement

Calibration metric

We summarize the semi-quantitative calibration metric introduced in Wagner-Carena et al. (2021). This is a multi-dimensional generalization of commonly used confidence-frequency tests (e.g. Niculescu-Mizil et al. 2005) for evaluating the statistical consistency of model uncertainties.

Denote the N parameter samples drawn from the Bayesian attentive neural process (BANP) posterior for some AGN k as $\{\boldsymbol{s}_n^{(k)}\}_{n=1}^N$ and the true parameter value as $\boldsymbol{s}_{\rm true}^{(k)}$. For a given fraction of the BANP posterior probability volume, p_X, the metric queries the fraction of the samples containing the truth within this volume, p_Y. More precisely,

$$p_Y^{(k)}(p_X) = \mathbb{1}\left\{\frac{\sum_{n=1}^N \mathbb{1}\left\{d(\boldsymbol{s}_n^{(k)}) < d(\boldsymbol{s}_{\rm true}^{(k)})\right\}}{N} < p_X\right\} \tag{6}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function that evaluates to 1 when the argument is true and 0 otherwise, and $d(\boldsymbol{s})$ is the distance measure of a particular point $\boldsymbol{s}$ from the posterior predictive mean given the posterior width. Following Park et al. (2021), we use the Mahalanobis distance for d. If the posterior is perfectly calibrated, p_X of the samples would encompass the truth $p_Y = p_X$ of the time, for every value value of p_X.

This metric can be evaluated on the dataset as a whole by averaging the p_Y values from individual AGN. For $N^{\rm val}$ AGN in the validation set, p_Y and p_X can be expressed as:

$$p_Y^{\rm val}(p_X) = \frac{1}{N^{\rm val}} \sum_{k=1}^{N^{\rm val}} p_Y^{(k)}(p_X) \tag{7}$$

Plotting $p_{\mathrm{Y}}^{\mathrm{val}}$ for a grid of p_X values yields the calibration curve. Regions of the curve with $p_{\mathrm{Y}}^{\mathrm{val}} < p_X$ indicate overconfidence whereas the opposite indicates underconfidence.