A Bayesian Perspective on the Deep Image Prior

Zezhou Cheng, Matheus Gadelha, Subhransu Maji, Daniel Sheldon

UMass Amherst

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

1. Revisiting the Deep Image Prior

➢ **Deep Image Prior (DIP)** [Ulyanov et al. 2018]

Random Noise | Generator | Corrupted Image | Learning | Restored Image (early stopping)

- A suitably designed deep network induces a natural image prior.
- Gradient descent over parameters to search over natural images.

➢ **Questions**

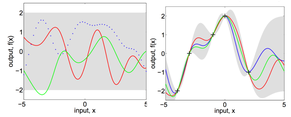
- Why does the DIP induce a natural image prior?
- Fitting parameters from a single image leads to overfitting. Earlier approach used early stopping which is hard to use in practice. Is there a principled approach to avoid overfitting?

➢ **Our contributions:**

- DIP is asymptotically equivalent to a stationary Gaussian process. We analytically derive the covariance for commonly used deep networks.
- Use SGLD to avoid overfitting and obtain a measure of uncertainty.

2. Revisiting Gaussian Process

➢ Gaussian Process (GP) defines a distribution over functions f(x).



GP Prior | GP Posterior

[images taken from Rasmussen 2006]

- Formally, $f(x) \sim \mathcal{GP}\big(m(x), K(x,x')\big)$ where $m(x) = \mathbb{E}\big(f(x)\big)$ and $\mathrm{K}(x,x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x')]$
- Any collection of function values have a joint Gaussian distribution, i.e., $[f(x_1), f(x_2), \ldots, f(x_n)]^T \sim \mathcal{N}(m, K)$

- Given $m = 0$, the posterior is $\boldsymbol{f}_B | X_B, X_A, \boldsymbol{f}_A \sim \mathcal{N}(K_{BA}K_{AA}^{-1}\boldsymbol{f}_A, K_{BB} - K_{BA}K_{AA}^{-1}K_{AB})$

References

[1] D. Ulyanov et al. Deep Image Prior. In CVPR-2018
[2] Rasmussen, and Williams. Gaussian Processes for Machine Learning. The MIT Press 2006.
[3] M. Welling et al. Bayesian Learning Via Stochastic Gradient Langevin Dynamics. In ICML-2011.

Source code and more results are available online.

3. Bayesian Interpretation of the Deep Image Prior

➢ **Theorem**

Let each channel of the input X be drawn independently from a zero mean stationary distribution with covariance function K_x. Then the output of a two-layer convolutional network with the sigmoid non-linearity, i.e., $h(t) = erf(t)$, converges to a zero mean stationary Gaussian process as the number of input channels c and filters H go to infinity sequentially. The stationary covariance K_z is given by

$$K_z^{erf}(t_1, t_2) = K_z^{erf}(r) = \frac{2}{\pi}\sin^{-1}\frac{K_x(r)}{K_x(0)}, \text{ where } r = t_2 - t_1$$

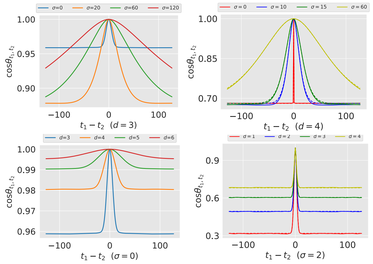
For a two-layer convolutional network with ReLU non-linearity,

$$K_z^{relu}(r) = \frac{K_x(0)}{2\pi}(\sin\theta_r^x + (\pi - \theta_r^x)\cos\theta_r^x), \text{ where } \cos\theta_r^x = \frac{K_x(r)}{K_x(0)}$$

- Assumption: the weights and biases are independent Gaussian RVs.

➢ **Beyond two layers**

- Bias term: $K_z^{bias}(r) = \sigma_b^2 + K_z(r)$
- Downsampling: $K_z^{\downarrow}(r) = K_z(\tau r)$
- Upsampling: $K_z^{\uparrow}(r) = K_z(r/\tau)$ (for band limited filter)
- Skip connections $K_z^{skip}(r) = \sum_i K_{z_i}(r)$ (for $z = \sum_i z_i$)



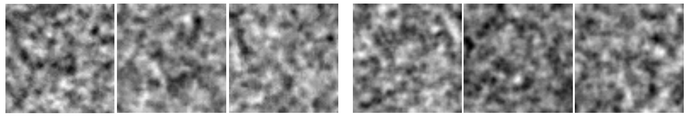
Auto-encoder | Conv-only network

➢ **Long-range spatial covariances**

Right figure: Covariance function $\cos\theta_t = K_z(t_1 - t_2)/K_z(0)$ w.r.t. different values of depth (d) and input covariance (σ).

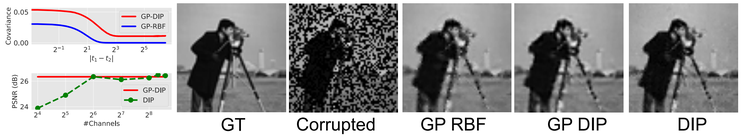
➢ **Experimental evidence**

- The samples drawn from the DIP and GP with equivalent stationary kernel.



DIP prior samples | GP prior samples

- The posterior mean estimated by the SGD with the DIP matches the GP posterior mean as the number of channels increases. However, posterior inference with GP kernels is **very slow** for large images compared to SGD.



GT | Corrupted | GP RBF (25.78 dB) | GP DIP (26.34 dB) | DIP (26.43 dB)

4. Bayesian Inference with the Deep Image Prior

➢ **SGLD** [M. Welling et al. 2011]

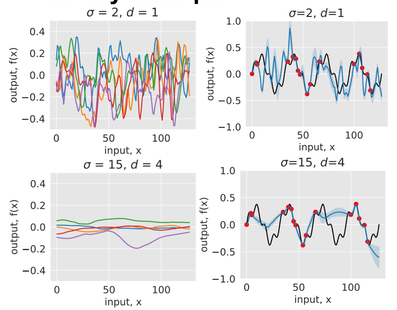
- Stochastic Gradient Langevin Dynamics
- Convert SGD into an MCMC sampler by injecting Gaussian noise to the gradient updates, formally,

$$\Delta_w = \frac{\epsilon}{2}(\nabla_w \log p(\hat{y}|w) + \nabla_w \log p(w)) + \eta_t$$

$$\eta_t \sim \mathcal{N}(0, \epsilon) \text{ where } \epsilon \text{ is step size}$$

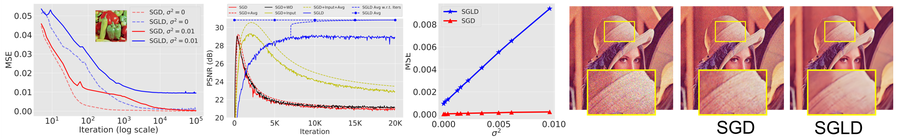
- Under suitable conditions (e.g. $\sum \epsilon_t = \infty$ and $\sum \epsilon_t^2 < \infty$ etc.), the samples from SGLD converges to the true posterior distribution.

➢ **1D Toy Example**



Prior | SGLD Posterior

➢ **Image Denoising**



MSE vs. Iteration | PSNR: SGD vs SGLD | MSE vs. noise var. | Corrupted | SGD (29.17 dB) | SGLD (32.05 dB)

	Hou.	Pep.	Lena	Bab.	F16	K1	K2	K3	K12	Avg.
SGD	26.74	28.42	29.17	23.50	29.76	26.61	28.68	30.07	29.78	28.08
SGLD	**30.86**	**30.82**	**32.05**	**24.54**	**32.90**	**27.96**	**32.05**	**33.29**	**32.79**	**30.81**

Image denoising task (PSNR)

➢ **Image Inpainting**

Corrupted | SGD (19.23 dB) | SGLD (21.86 dB) | Uncertainty

- ❑ SGLD enables us estimate the variance from posterior samples
- ❑ SGLD consistently outperforms vanilla gradient descent on image denoising/inpainting

	Barb.	Boat	Hou.	Lena	Pep.	C.m.	Cou.	Fin.	Hill	Man	Mon.	Avg.
SGD	28.48	31.54	35.34	35.00	30.40	27.05	30.55	32.24	31.37	31.32	30.21	28.08
SGLD	**33.82**	**34.26**	**40.13**	**37.73**	**33.97**	**30.33**	**33.72**	**33.41**	**34.03**	**33.54**	**34.65**	**34.51**

Image inpainting task (PSNR)