Amortised MAP Inference for Image Super-resolution

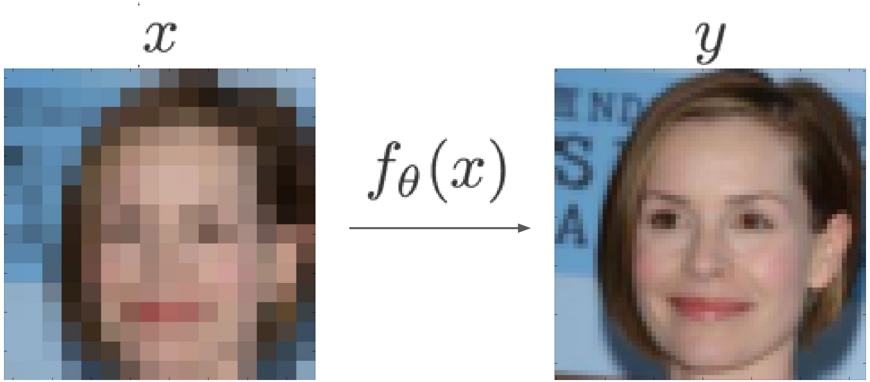
Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi & Ferenc Huszár

ICLR 2017

+ MAGIC PONY TECHNOLOGY

Super Resolution

- Inverse problem: Given low resolution representation x reconstruct high resolution image y



Ranked Inference Choices

1. Empirical Risk Minimization: minimize a loss function measuring what we care about
   (We don't know the right loss function, in fact we can't even measure perceptual quality)

$$\min_{\theta} \mathbb{E}_{y,x}[\ell(y, f_\theta(x))] = \min_{\theta} \mathbb{E}_{y,x}[\ell(y, \hat{y})]$$

2. Maximum a Posteriori (MAP) inference using knowledge of image prior
   (We don't know the prior)

3. Approximate MAP

Motivation: Blurry Images

- MSE is the wrong objective for photo-realistic results

MSE Super-Resolution (4x)

Original HR

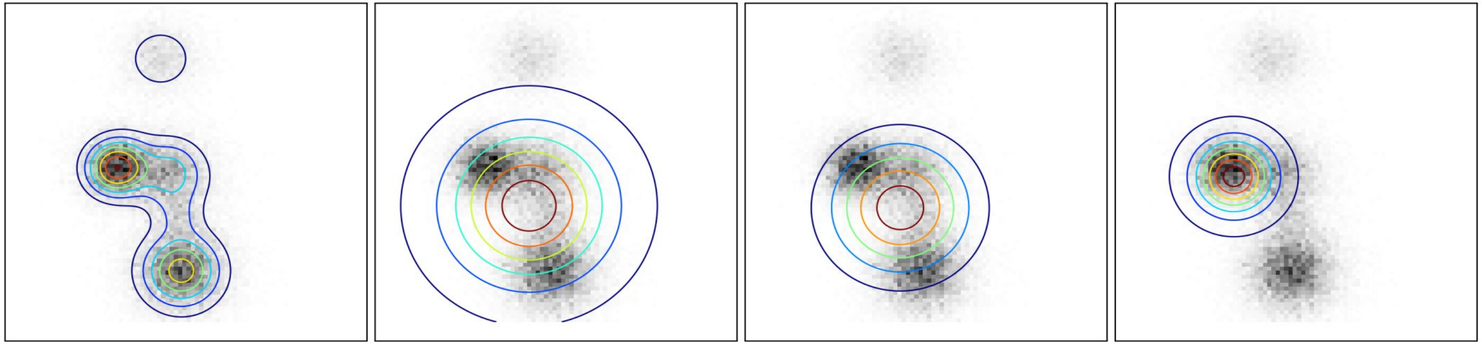
Motivation: The Divergence Perspective

Data

KL[P|Q]
Maximum Likelihood
(overdispersed)

JSD[PIQ]
Original GAN criterion
(in between)

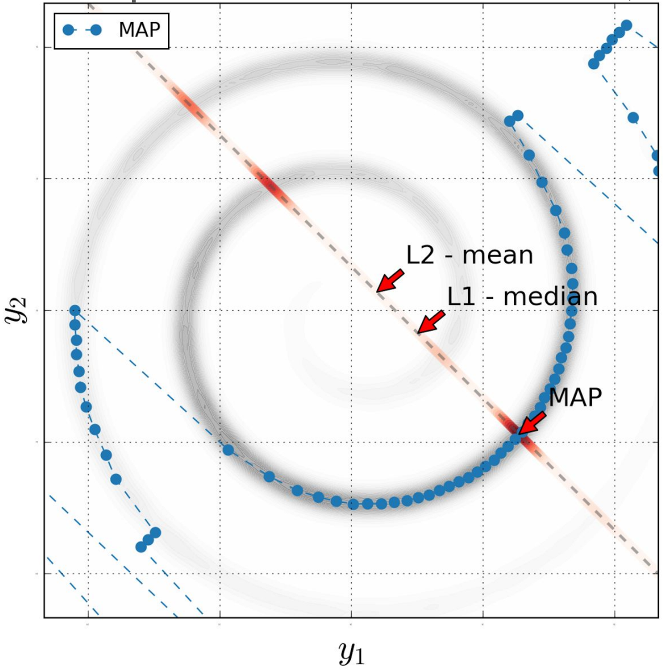
KL[QIP]
Modified GAN criterion
(mode seeking)



Motivation: The 2D Perspective

Target: 2D High resolution, $y = [y_1, y_2]$ drawn from a Swiss-roll

Input: 1D Low resolution, $x = (y_1+y_2)/2$, average of high resolution



Example

Input: x = 0.5

Valid outputs fall on the line: $y_1 = 1 - y_2$

Approximate Amortized MAP Inference

- Maximize Log-posterior evaluated at the predicted output

$$\underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{x\sim p_x} \log p_{Y|X}(\underbrace{f_\theta(x)}_{\hat{y}}|x)$$

- Decomposed via Bayes' rule

$$\underset{\theta}{\operatorname{argmax}} \left\{ \underbrace{\mathbb{E}_x \log p_{X|Y}(x|f_\theta(x))}_{\text{Likelihood}} + \underbrace{\mathbb{E}_x \log p_Y(f_\theta(x))}_{\text{Prior}} \right\}$$

Likelihood
data consistency

Prior
plausible output

Data Consistency: affine projections

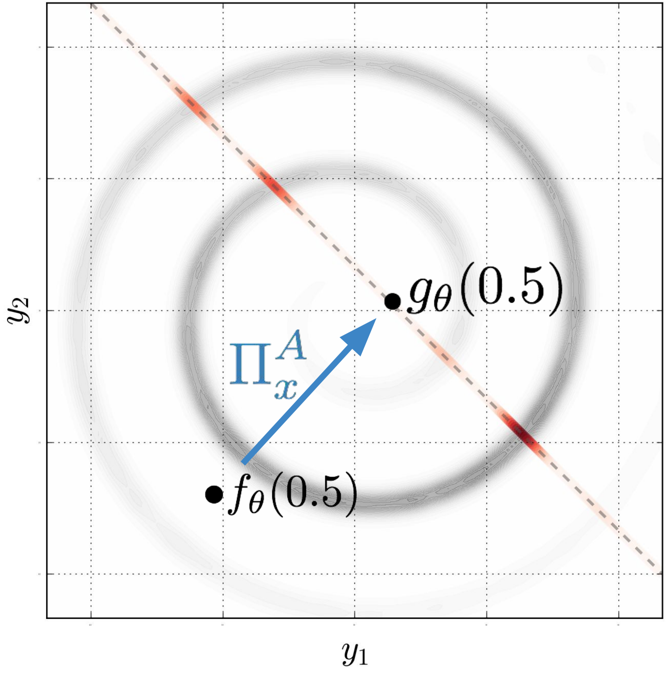
- Input $x = Ay$

  $\underbrace{A}$ Linear downsampling (strided convolution)

- Affine Projected Network

$$g_\theta(x) = \Pi_x^A f_\theta(x)$$
$$= (I - A^+ A) f_\theta(x) + A^+ x$$

  $\underbrace{A^+}$ Moore-Penrose pseudo inverse of A (subpixel (transposed) convolution)



- For Affine projected networks the likelihood is always maximally satisfied and can be ignored

$$\underset{\theta}{\operatorname{argmax}} \{ \underbrace{\mathbb{E}_x \log p_{X|Y}(x|g_\theta(x))} + \underbrace{\mathbb{E}_x \log p_Y(g_\theta(x))} \}$$

Plausible output: Cross Entropy

- Cross Entropy objective

$$\underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{x\sim p_X} \log p_Y\left(g_\theta(x)\right) = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{\hat{y}\sim q_\theta} \log p_Y\left(\hat{y}\right)$$

$$= \underset{\theta}{\operatorname{argmin}}\, \mathbb{H}[q_\theta, p_Y], \quad \forall x: \;\; Ag_\theta(x) = x$$

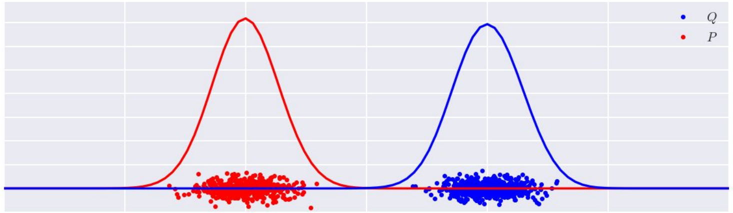
- We propose three methods to optimize this
  1. AffGAN: GAN using modified update rule minimizing $\mathrm{KL}[q_\theta \| p_Y] = \mathbb{H}[q_\theta, p_Y] - \mathbb{H}[q_\theta]$
  2. AffDAE: Optimize using gradients from denoiser $\frac{\partial \log p(y)}{\partial y} = \frac{f^*_\sigma(\tilde{y}) - y}{\sigma^2} + \mathcal{O}(\sigma^2) \quad \text{as} \quad \sigma \to 0$
  3. AffLL: Fit differentiable parametric density model to the image distribution and get gradient estimates directly.

Fix GAN Instability: Instance Noise

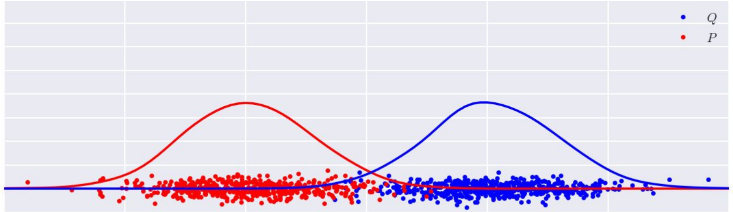
- The data and model distributions will most likely not share support
  - Discriminator can always perfectly separate the samples
  - KL-divergence is undefined and in general the GAN convergence proof does not hold
  - No gradient to train generator
- Instance Noise: Broaden support of model and data distributions by adding gaussian noise
  - Minimises: $\underset{\theta}{\operatorname{argmin}} \operatorname{KL}\left[q_\theta * p_n || p_Y * p_n\right]$

(Arjovsky & and Bottou, ICLR 2017 arrives at the same solution)

Perfectly separable distributions



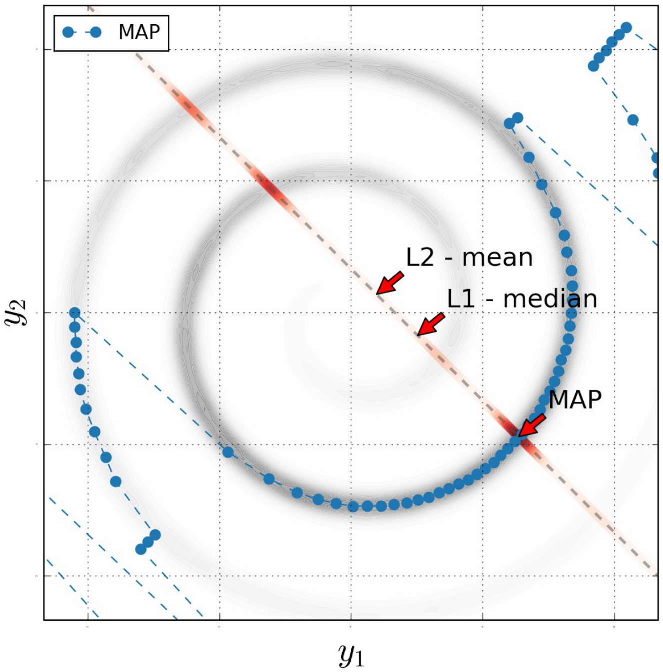
Instance Noise

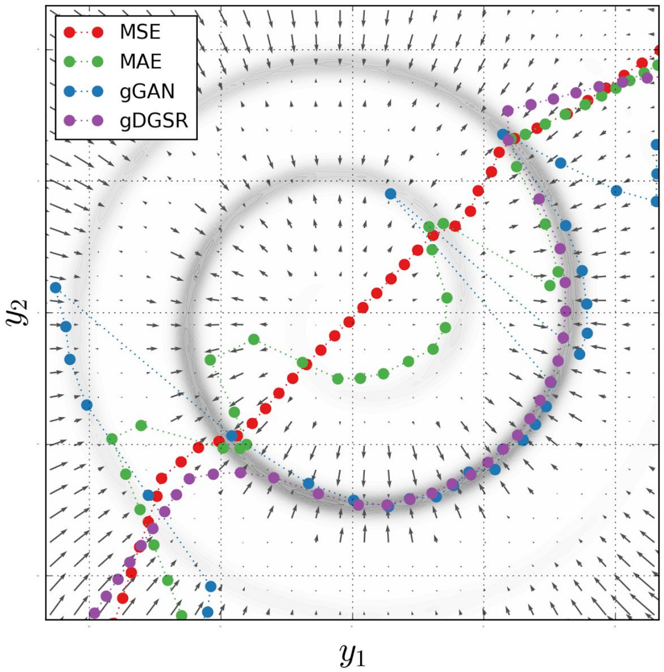


2D Super resolution - Revisited

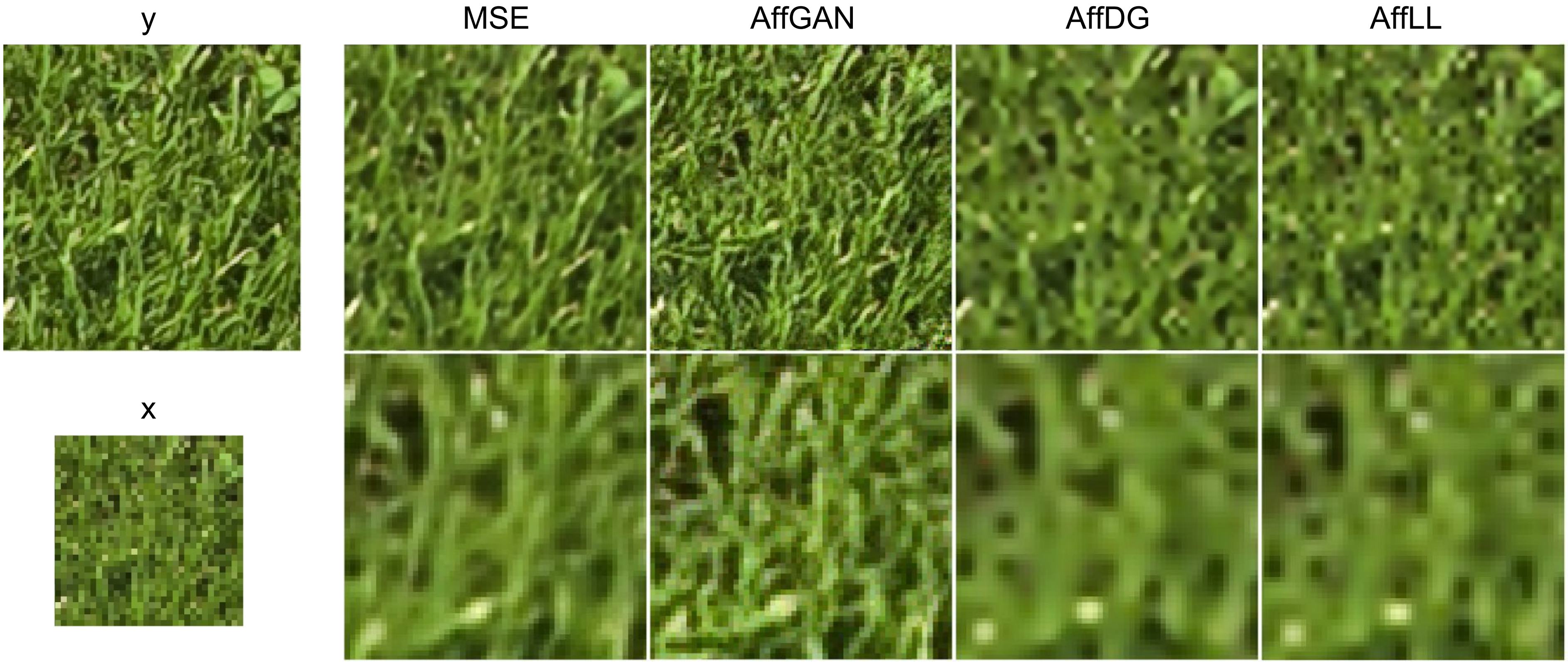
Target: 2D High resolution, $y = [y_1, y_2]$ drawn from a Swiss-roll

Input: 1D Low resolution, $x = (y_1+y_2)/2$, average of high resolution





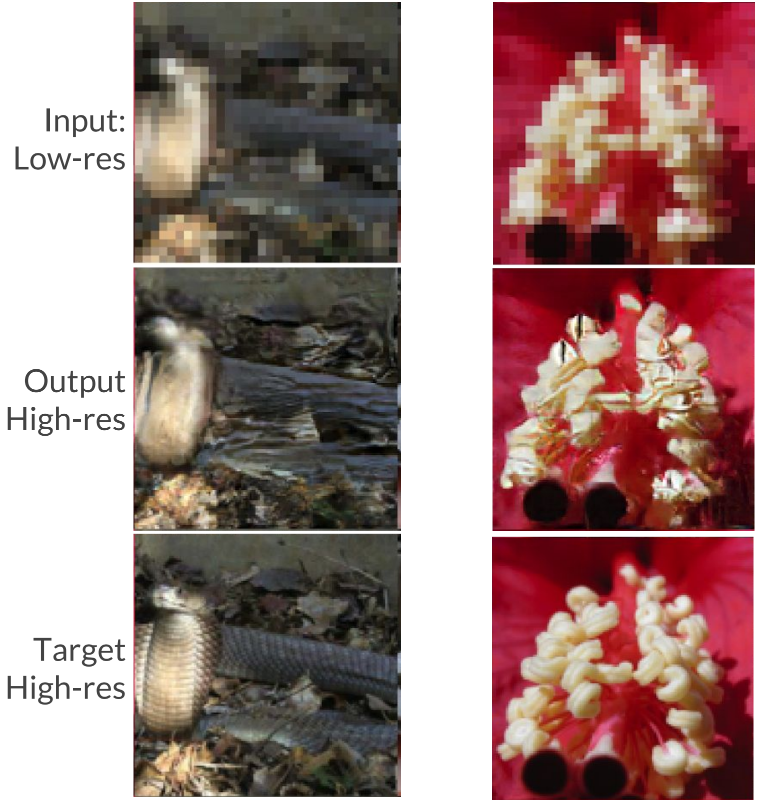
Results - Grass textures



Natural Images

- Mode seeking behavior

- Correct output is often ambiguous
  - Stochastic extension



AffGAN as Variational Inference

- Stochastic Affine Projected GAN (StAffGAN)
  (All two letter GAN acronyms were occupied)

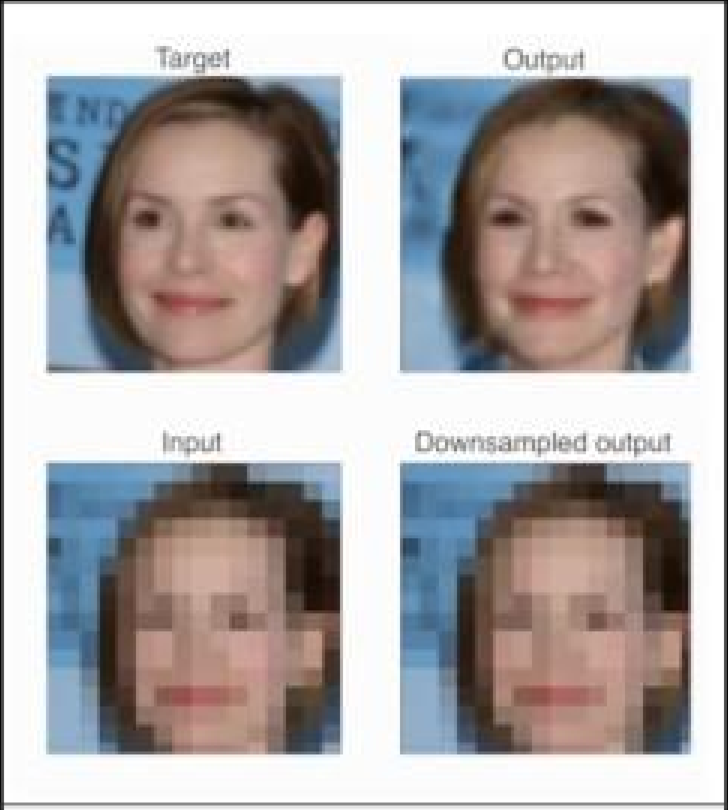
- Add noise source z in addition to input x

$$q_{Y;\theta} = \mathbb{E}_{x \sim p_X} \mathbb{E}_{z \sim p_Z} \delta\left(y - g_\theta(x, z)\right)$$

- For super-resolution we can show that

$$\operatorname*{argmin}_\theta \mathrm{KL}[q_{Y;\theta} \| p_Y] = \operatorname*{argmin}_\theta \mathrm{KL}[q_{Y|X;\theta} \| p_{Y|X}]$$

- Consequence: StAffGAN performs **Variational Inference** for image super-resolution only requiring samples from p_X and p_y



Target
Output
Input
Downsampled output


AffGAN
BS-GAN
Conditional GAN
DCGAN
EBGAN
f-GAN
GAN
InfoGAN
JS-GAN
LSGAN
Mode-seeking GAN
Temporal GAN
GRAN
StAffGAN
Unrolled GAN
VEGAN
VGAN
Wasserstein GAN
SRGAN

Conclusion

- Affine projections restricting model to the affine subspace of valid solutions

- We proposed three methods for amortized MAP inference in image super-resolution
  - A modified GAN objective
  - Denoiser guided optimization
  - Direct parametric likelihood modelling

- In practice the GAN based objective produced the visually most appealing results

- Provide theoretical grounding for GANs in generative models

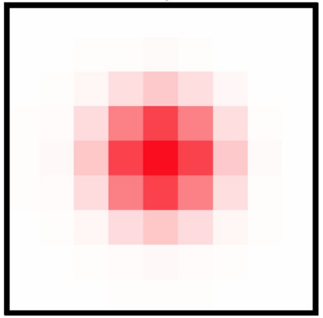
- We showed how GAN models be seen as performing Amortized Variational Inference

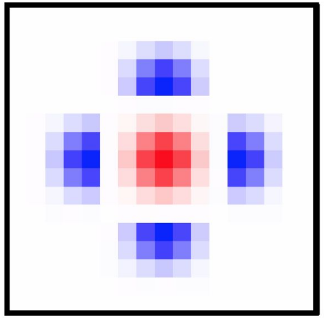
Further Slides

How to compute A^+

- In the language of Deep Learning:
  - A is a strided convolution with, say, Gaussian kernel
  - A^+ is a subpixel (transposed) convolution
- For fixed A, we find A^+ via numerical optimization:

$$\ell_1(B) = \mathbb{E}_{y\sim\mathcal{N}_{rd}} \|Ay - ABAy\|_2^2$$
$$\ell_2(B) = \mathbb{E}_{x\sim\mathcal{N}_{d}} \|Bx - BABx\|_2^2$$
$$A^+ = \text{argmin}_B(\ell_1(B) + \ell_2(B))$$

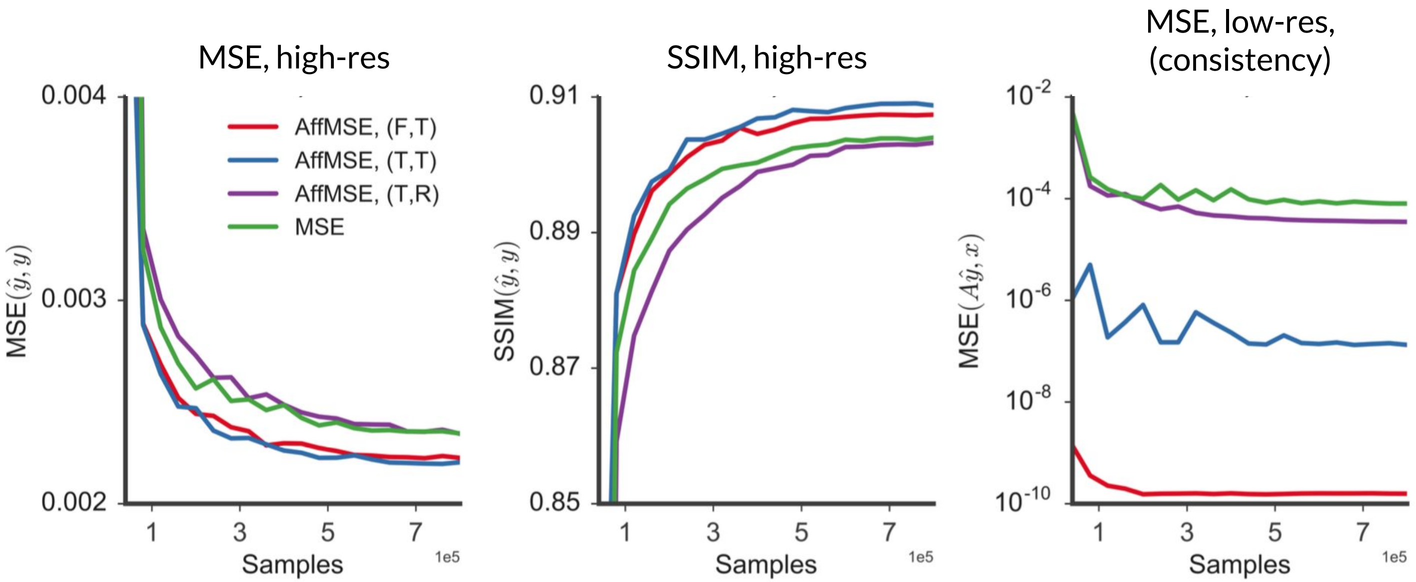




$$\max_\theta \left\{ \underbrace{\mathbb{E}_x \log p(x|\hat{y})}_{\text{Likelihood}} + \underbrace{\mathbb{E}_x \log p(\hat{y})}_{\text{Prior}} - \underbrace{\mathbb{E}_x \log p(x)}_{\text{Evidence}} \right\}$$

Can Affine Projected Networks learn?

- Proof of Concept using MSE



Legend tuple indicate:
(**F**)ixed / (**T**)rainable affine transformation, (**R**)andom / (**T**)rained initialization of transformation

Approximate MAP inference - Method 1: Model p(y)

- Directly model log[p(y)] using maximum likelihood learning
  - Fit differentiable parametric model to log[p(y)]
  - maximize f(x) using gradients of log[p(y)]

- We use with PixelCNN + MCGSM
  - PixelCNN: Convolutional Network satisfying the Chain Rule*
  - MCGSM: Mixture of Conditional Gaussians Scale Mixture model**

*Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759, 2016.
** Theis, Lucas, and Matthias Bethge. Generative image modeling using spatial lstms. Advances in Neural Information Processing Systems. 2015.

$$\max_{\theta}\left\{\underbrace{\mathbb{E}_x \log p(x|\hat{y})}_{\text{Likelihood}} + \underbrace{\mathbb{E}_x \log p(\hat{y})}_{\text{Prior}} - \underbrace{\mathbb{E}_x \log p(x)}_{\text{Evidence}}\right\}$$

Approximate MAP inference - Method 2: DAE

- To maximize the prior we need the gradients

$$\frac{\partial}{\partial \theta}\mathbb{E}_x[\log p(\hat{y})] = \mathbb{E}_x\left[\frac{\partial}{\partial y}\log p(y)\cdot\frac{\partial}{\partial \theta}\hat{y}\right]$$

- Luckily these can be approximated using a Denoising Autoencoder *

$$\frac{\partial \log p(y)}{\partial y} = \frac{f_\sigma^*(\tilde{y}) - y}{\sigma^2} + \mathcal{O}(\sigma^2) \quad \text{as} \quad \sigma \to 0$$

$$\max_\theta \left\{ \underbrace{\mathbb{E}_x \log p(x|\hat{y})}_{\text{Likelihood}} + \underbrace{\mathbb{E}_x \log p(\hat{y})}_{\text{Prior}} - \underbrace{\mathbb{E}_x \log p(x)}_{\text{Evidence}} \right\}$$

*Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. Journal of Machine Learning Research, 15(1):3563–3593, 2014.

Approximate MAP inference - Method 3: GAN

- We propose a modified GAN update rule that minimizes KL[Q|P]

$$\begin{aligned} KL[Q|P] &= \mathbb{E}_{y\sim Q}\left[\log Q(y) - \log P(y))\right] \\ &= \mathbb{E}_{y\sim Q}\left[\log Q(y)\right] - \mathbb{E}_{y\sim Q}\left[\log P(y)\right] \\ &= -H_y(Q) - \mathbb{E}_{y\sim Q}\left[\log P(y)\right] \end{aligned}$$

- Minimizing KL[Q|P] ⇒ Maximize MAP + Entropy-term

$$\max_{\theta}\left\{\underbrace{\mathbb{E}_x \log p(x|\hat{y})}_{\text{Likelihood}} + \underbrace{\mathbb{E}_x \log p(\hat{y})}_{\text{Prior}} - \underbrace{\mathbb{E}_x \log p(x)}_{\text{Evidence}}\right\}$$

Super Resolution

- Inverse problem: Given low resolution representation x reconstruct high resolution image y

- Ambiguous / Multimodal