UNIVERSITY OF CAMBRIDGE · HERIOT WATT UNIVERSITY · INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR · UNIVERSITY OF BIRMINGHAM

Unsupervised Training of Convex Regularizers using Maximum Likelihood Estimation

Hong Ye Tan

Joint work with: Ziruo Cai, Marcelo Pereyra, Subhadip Mukherjee, Junqi Tang, Carola-Bibiane Schönlieb

Department of Applied Mathematics and Theoretical Physics, University of Cambridge

Maximum Marginal Likelihood Estimation

- **Goal:** learning a regularizer from only measurements
- **Setting:** one-shot corrupted dataset, no ground truth
  - Rules out Noise2Noise, Noise2Inverse
  - Still OK: Equivariant Imaging, deep image prior
  - Non-blind: assume known forward operator (i.e. likelihood)
- **Existing Bayesian methods**: hand-crafted models (e.g. TV, wavelet), one image at a time
- **This**: neural network regularizer, for a whole dataset

Method

Bayesian approach: maximum likelihood estimation.

- Given only measurements y, find the best θ that fits the data
- For a data prior $p(x|\theta) \propto \exp(-g_\theta(x))$ and likelihood $\ell(y|x)$,

MMLE: ★ $\boxed{\theta^* = \arg\max_{\theta\in\Theta} \log p(y|\theta)}$ $\quad p(x|\theta) \propto \exp(-g_\theta(x))$, $\quad p(y|\theta) = \int \ell(y|x)p(x|\theta)\mathrm{d}x$

(Regularity conditions) ⇓ Want to solve with gradient methods

Fisher's Identity $\quad \nabla_\theta \log p(y|\theta) = \mathbb{E}_{x|\theta}[\nabla_\theta g_\theta(x)] - \mathbb{E}_{x|y,\theta}[\nabla_\theta g_\theta(x)]$

Main Idea

Use the current θ to sample from $p(x|\theta)$, $p(x|y,\theta)$

⇓ Use MCMC to approximate expectations: unadjusted Langevin algorithm (ULA)

sampling from:

$$\mathrm{R}_{\gamma,\theta}: \; X_{k+1} = X_k - \gamma\nabla_x(f_y + g_\theta)(X_k) + \sqrt{2\gamma}Z_{k+1} \qquad p(x|y,\theta)$$

$$\bar{\mathrm{R}}_{\gamma',\theta}: \; \bar{X}_{k+1} = \bar{X}_k - \gamma'\nabla_x g_\theta(\bar{X}_k) + \sqrt{2\gamma'}Z'_{k+1} \qquad p(x|\theta)$$

⇓ Apply estimate for gradient ascent on θ

$$\theta_{n+1} = \theta_n + \frac{\delta_{n+1}}{m_n}\sum_{k=1}^{m_n}\{\underbrace{\nabla_\theta g_\theta(\bar{X}_k^n)}_{\approx\, \mathbb{E}_{x|y,\theta}[\nabla_\theta g_\theta(x)]} - \underbrace{\nabla_\theta g_\theta(X_k^n)}_{\approx\, \mathbb{E}_{x|\theta}[\nabla_\theta g_\theta(x)]}\} \qquad \text{(SAPG)}$$

★ **Computational efficiency:** $m_n = 1$ is enough! [2]

Theorem. Assume that $-\log p(y|\theta)$ is convex w.r.t. θ. If

$$g_\theta : \mathbf{x} \mapsto \sum_{i=1}^{C} \psi_i(\mathbf{w}_i^\top \mathbf{x})$$

takes the form of a convex ridge regularizer, then the SAPG iterates converge ergodically to the maximum marginal likelihood estimator.

Reconstruction:

$y = Ax + \varepsilon \quad$ Fidelity $f_y(x)$, regularizer $g_\theta(x)$

Variational formulation		MAP estimation
$\hat{x} = \arg\min_x [f_y(x) + g_\theta(x)]$	$\xleftrightarrow[g_\theta(x) = -\log p(x\vert\theta)]{f_y(x) = -\log \ell(y\vert x)}$	$x_{\mathrm{MAP}} = \arg\max_x \log[\ell(y\vert x)p(x\vert\theta)] = \arg\max_x \log p(x\vert y,\theta)$

&

$x_{\mathrm{MMSE}} = \mathbb{E}[x|y,\theta]$ using MCMC

MMSE estimation

Convex Ridge Regularizer[3]

$$g_\theta : \mathbf{x} \mapsto \sum_{i=1}^{C} \psi_i(\mathbf{w}_i^\top \mathbf{x})$$

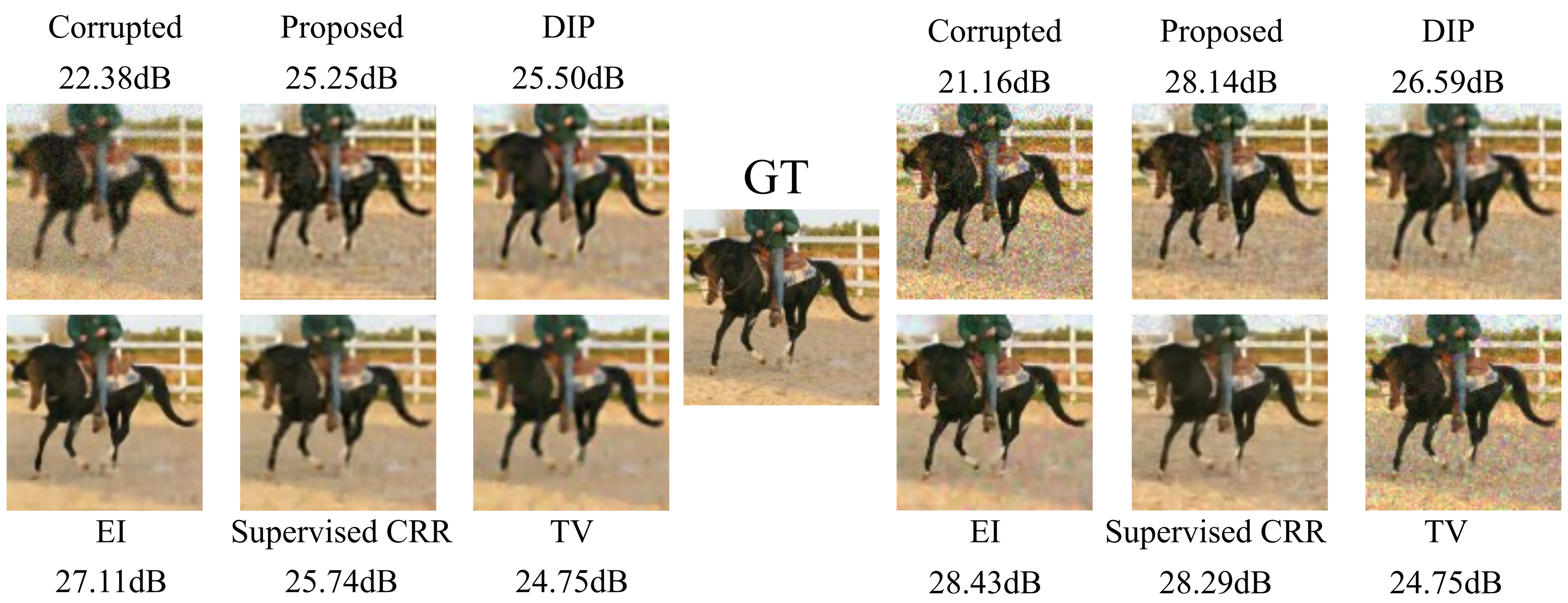
ψ_i convex "profile" functions: piecewise quadratic splines

$\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_C]$ parameterized as convolution

- Convergent MAP estimation
- "Medium" parameter # (~ 10^5)
- Supervised version: unrolled gradient step
- Easy to compute derivative (no autograd)

Experiments

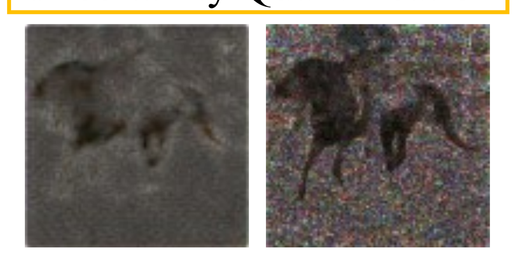
Gaussian deconvolution | STL-10 | Poisson denoising



Deep image prior (**DIP**)
- Corrupted measurement only
- Not applicable to datasets
- Heuristic early stopping

Equivariant Imaging (**EI**)
- Unsupervised (same setting)
- Requires knowledge of forward operator group invariances
- End-to-end (~10^7 parameters)

Uncertainty Quantification



+ Working high-dimensional Bayesian optimization
+ Fast optimization-based MAP estimation
+ Uncertainty quantification for MMSE estimation
+ Model generalization to different forward operators

- Slow training (~3 days instead of ~3 hours for supervised)
- Monte Carlo: slow convergence of MMSE estimates (~20k samples)
- Strong convergence assumption on data

Key Takeaways:

1. Working application of stochastic optimization in high dimensions (CRR network parameters, from $\dim\theta \sim 10^1$ to $\sim 10^5$)
2. Small performance gap to full supervision with same architecture ★
3. Leverage statistics of many corrupted images to create a strong prior

References

[1] HYT, ZC, MP, SM, JT, CBS. Unsupervised Training of Convex Regularizers using Maximum Likelihood Estimation. TMLR 2024.

[2] Vidal, De Bortoli, Pereyra, Durmus. Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: an empirical Bayesian approach. SIAM IS 2020.

[3] Goujon, Neumayer, Bohra, Ducotterd, Unser. A neural network based convex regularizer for inverse problems. IEEE TCI 2023.

Paper (TMLR'24)