Learning convex regularizers satisfying the variational source condition for inverse problems

Subhadip Mukherjee[1], **Carola-Bibiane Schönlieb**[1], **and Martin Burger**[2]
[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
[2]Department of Mathematics, University of Erlangen-Nuremberg, Germany
Emails: {sm2467,cbs31}@cam.ac.uk, martin.burger@fau.de

Abstract

Variational regularization has remained one of the most successful approaches for reconstruction in imaging inverse problems for several decades. With the emergence and remarkable success of deep learning in recent years, a considerable amount of research has gone into data-driven modeling of the regularizer in the variational setting. Our work extends a recently proposed method, referred to as *adversarial convex regularization* (ACR), that seeks to learn a data-driven convex regularizer via adversarial training in an attempt to combine the power of data with the classical convex regularization theory. Specifically, we leverage the variational source condition (SC) during training to enforce that the ground-truth images minimize the variational loss corresponding to the learned convex regularizer. This is achieved by adding an appropriate penalty term to the ACR training objective. The resulting regularizer (abbreviated as ACR-SC) performs on par with standard ACR, but unlike ACR, comes with a quantitative convergence rate estimate.

1 Introduction

Linear inverse problems seek to recover an unknown parameter $x^* \in \mathcal{X}$ from its noisy measurement $y^\delta = Ax^* + e \in \mathcal{Y}$, where $A : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator between the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, and e denotes measurement noise with $\|e\|_{\mathcal{Y}} \leq \delta$. The clean measurement corresponds to $\delta = 0$ and is denoted by y^0. Inverse problems are typically ill-posed, in the sense that A is either non-invertible or poorly conditioned, leading to noise amplification in the solution obtained via naïve inversion. Variational methods [4, 13], wherein one seeks to trade-off data-fidelity with a regularizer, has traditionally been the most popular and successful approach for computing a stable solution to ill-posed inverse problems, and are rooted in a rigorous function-analytic foundation.

With the advent of deep learning, considerable attention has been devoted to leveraging the availability of data for solving inverse problems [3]. A particularly notable deep learning-based endeavor has sought to model the regularizer via over-parametrized deep neural networks and learn it in a data-driven manner, instead of using hand-crafted functionals [9, 10, 12, 8]. Such methods inherit the theoretical guarantees (such as stability and convergence) offered by the variational framework while effectively utilizing the power of data, provided that the regularizer fulfills certain conditions (such as convexity [12]). This work essentially builds upon the approach introduced in [12]. By incorporating an additional penalty term during training, our approach encourages the ground-truth images to satisfy the so-called variational source condition (SC), which leads to precise convergence rate estimates.

2 Variational source condition

In [5], convergence rate estimates for variational reconstruction with convex regularizers were derived under the source condition, which provides an additional regularity condition on the solution of

NeurIPS 2021 Workshop on Deep Learning and Inverse Problems, virtual.

ill-posed inverse problems (see [4] for a detailed discussion). Here, we briefly recall the key results to make the exposition self-contained. Consider the variational regularization approach that minimizes an energy functional, given by

$$x_\lambda \in \arg\min_{x\in\mathcal{X}} \frac{1}{2}\|y^\delta - Ax\|_2^2 + \lambda\,\psi_\theta(x). \tag{1}$$

Here, $\{\psi_\theta\}_{\theta\in\Theta}$ is a convex regularizer parametrized by an input-convex neural network (ICNN) [2].

Definition 1 (Source condition)**.** *The ψ_θ-minimizing solution is defined as*

$$\tilde{x} \in \arg\min_{x\in\mathcal{X}} \psi_\theta(x) \text{ subject to } Ax = y^0. \tag{2}$$

The variational source condition is satisfied if there exists some $\tilde{w} \in \mathcal{Y}$ such that $A^*\tilde{w} \in \partial\,\psi_\theta(\tilde{x})$.

One can show that the set of $\tilde{x}$ satisfying the SC is the same as the set of solutions to the variational problem (1), see [5, Proposition 1]. This serves as the main motivation behind our approach, in that we encourage the ground-truth images during training to satisfy the SC and thereby be the solution to the variational problem corresponding to the learned regularizer.

Let $D_{\psi_\theta}(x_1, x_2) := \left\{\psi_\theta(x_1) - \psi_\theta(x_2) - \langle u, x_1 - x_2\rangle \middle| u \in \partial\psi_\theta(x_2)\right\}$ be the Bregman distance corresponding to ψ_θ. Then, the following convergence rate can be established (see [5]):

Theorem 1. *Let $\|y^\delta - y^0\| \le \delta$, $\tilde{x}$ be a ψ_θ-minimizing solution as defined in* (2)*, and suppose that the SC holds. Then, for each minimizer x_λ of* (1)*, there exists $d \in D_{\psi_\theta}(x_\lambda, \tilde{x})$ such that $d \le \lambda\frac{\|\tilde{w}\|^2}{2} + \frac{\delta^2}{2\,\lambda}$ holds. Therefore, choosing $\lambda \propto \delta$ leads to an $\mathcal{O}(\delta)$ convergence rate.*

Consequently, SC enables one to derive quantitative convergence rate estimates in terms of the Bregman distance induced by the learned convex regularizer. In contrast, [12] shows convergence, without any quantitative rate estimate, while having to incorporate an additional Tikhonov term.

2.1 Learning ACR with the source condition (ACR-SC)

If the forward operator A is invertible, SC dictates that $\ell_{\text{sc}}(x;\theta) = \|(A^*)^{-1}\nabla_x\psi(x;\theta)\| < \infty$ must be satisfied whenever x is a solution to (1). Therefore, the smaller the quantity $\ell_{\text{sc}}(x;\theta)$ is, the more suitable x would be as a solution to (1). A natural way to encourage this is to penalize $L_{\text{sc}}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_{\text{sc}}(x_i;\theta)$, where $x_i \sim \mathbb{P}_r$ are the ground-truth images. For non-invertible A, one can replace the inverse with the Moore-Penrose pseudo-inverse. Denoting by $\mathbb{P}_n$ the distribution of the undesirable images in the adversarial training framework, the overall training objective becomes

$$L(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_\theta(x_i) - \frac{1}{n}\sum_{i=1}^n \psi_\theta(z_i) + \lambda_{\text{gp}}L_{\text{gp}}(\theta) + \lambda_{\text{sc}}L_{\text{sc}}(\theta). \tag{3}$$

Here, $x_i \sim \mathbb{P}_r$ and $z_i \sim \mathbb{P}_n$ are the ground-truth and noisy/undesirable images in the training dataset, respectively. The gradient penalty is given by $L_{\text{gp}}(\theta) = \frac{1}{n}\sum_{i=1}^n \left(\|\nabla\psi_\theta\left(\epsilon x_i + (1-\epsilon)z_i\right)\|_2 - 1\right)^2$, where $\epsilon \sim$ uniform$[0,1]$, which encourages the regularizer to be 1-Lipschitz (cf. [10, 6]). Notably, the training loss in (3) can be computed without direct supervision (i.e., without pairs of noisy and ground-truth images), thus offering more flexibility. For solving the variational problem (1) with the learned convex ψ_θ, one can employ a simple iterative sub-gradient algorithm for minimization. Notably, an advantage of the enforced SC is that Bregman iteration techniques are available for (1) and well understood (cf. [4] for details). We show some preliminary promising results here (see Figure 1) and will further pursue this approach in future research.

3 Numerical results

We first show a proof-of-concept of the proposed ACR-SC method considering a denoising problem on MNIST and subsequently compare it with a number of state-of-the-art supervised and unsupervised data-driven methods for the prototypical inverse problem of computed tomography (CT) reconstruction from sparse-view projection data. In both experiments, the reconstruction quality is evaluated in terms of the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) with respect to the target ground-truth.

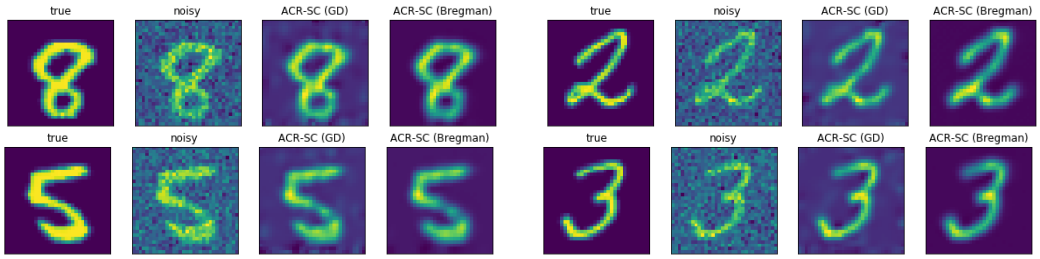


Figure 1: Representative denoising examples on MNIST. The average PSNR and SSIM over 100 randomly chosen test images are as follows: (i) noisy: 13.93 ± 0.13 dB, 0.51 ± 0.08 ; (ii) ACR-SC (GD): 22.72 ± 0.64, 0.77 ± 0.04; and (iii) ACR-SC (Bregman): 20.29 ± 0.88, 0.86 ± 0.03. The Bregman technique is executed with $\lambda = 25$ as opposed to $\lambda = 5$ in vanilla GD, and it does a comparatively better job of recovering the contrast while yielding effective denoising.

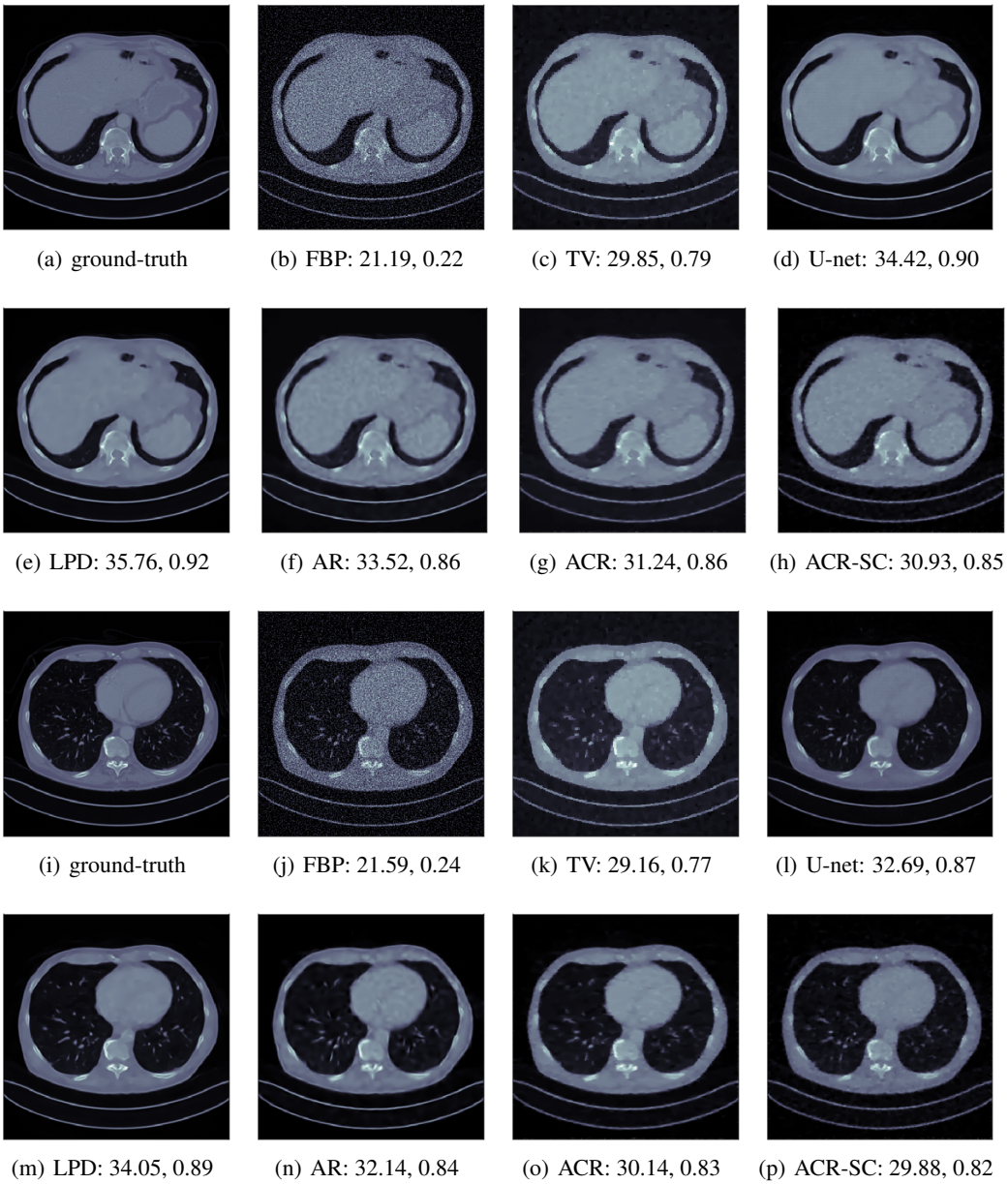
method	**PSNR (dB)**	**SSIM**	**# param.**	**reconstruction time (ms)**
FBP	21.28 ± 0.13	0.20 ± 0.02	1	37.0 ± 4.6
TV	30.31 ± 0.52	0.78 ± 0.01	1	28371.4 ± 1281.5
Supervised methods				
U-Net	34.50 ± 0.65	0.90 ± 0.01	7215233	44.4 ± 12.5
LPD	35.69 ± 0.60	0.91 ± 0.01	1138720	279.8 ± 12.8
Unsupervised methods				
AR	33.84 ± 0.63	0.86 ± 0.01	19338465	22567.1 ± 309.7
ACR	31.55 ± 0.54	0.85 ± 0.01	606610	109952.4 ± 497.8
ACR-SC	31.28 ± 0.50	0.84 ± 0.01	590928	105232.1 ± 378.5

Table 1: PSNR and SSIM statistics over 128 test slices for CT reconstruction on Mayo-clinic data. ACR-SC is competitive with ACR, uses fewer learnable parameters, and leads to convergence rate estimates.

For the denoising experiment on MNIST, the clean digits are corrupted with an additive white Gaussian noise with $\sigma = 0.2$. The regularizer is trained over 10 epochs with a batch-size of 64 (with $\lambda_{\text{sc}} = 2.0$ and $\lambda_{\text{gp}} = 10.0$), by minimizing the training loss (3). Subsequently, the learned regularizer is plugged into the variational framework (1). We employ both gradient-descent (with $\lambda = 5$, 300 iterations with a step-size of 0.01) and Bregman iterations [4] (with $\lambda = 25$) for solving (1). The numerical examples shown in Fig. 1 demonstrate that ACR-SC performs a reasonable denoising and significantly improves the PSNR and SSIM over the noisy input. As shown in Fig. 1, Bregman iterations do a noticeably better job of circumventing loss of contrast in the reconstruction.

For Sparse-view CT reconstruction, We adopt the same experimental setting considered in [12] to ensure a fair comparison, and we briefly recall it here for the sake of completeness. For training the models, we use the publicly available data for the Mayo-clinic low-dose CT challenge [11]. All models are trained on 9 patients (2250 2D slices) and tested on the remaining one patient (128 slices). Parallel-beam projection data are simulated in ODL [1] with 200 uniformly spaced angles and 400 lines per angle, followed by additive Gaussian noise contamination (with $\sigma = 2.0$). For comparison, we consider the classical filtered back-projection (FBP) and total variation (TV) regularization as two representative model-based approaches. Among data-driven approaches, we compare with two supervised methods, namely (i) U-Net-based post-processing of FBP [7] and (ii) learned primal-dual (LPD); and two unsupervised approaches, namely (i) the adversarial regularization (AR) method introduced in [10] and (ii) its convex variant ACR [12]. Together with PSNR and SSIM, the reconstruction times are also reported to facilitate an easier assessment of the quality-vs.-time trade-off for different approaches (c.f. Table 1).

The architecture of the ICNN that models the regularizer in ACR-SC is taken to be identical to the one considered in [12]. However, unlike [12], our regularizer does not have the sparsifying filter-bank and the squared-ℓ_2 terms. We choose $\lambda_{\text{sc}} = 2.0$ and $\lambda_{\text{gp}} = 5.0$, and the regularizer is trained over 10 epochs with a batch-size of four. Adam optimizer with $\eta, \beta_1, \beta_2 = 10^{-5}, 0.90, 0.99$ is used for training. The variational problem with the resulting regularizer is solved via gradient-descent ($\lambda = 0.05$, step-size 0.8, 400 iterations). The numerical results reported in Table 1 and Fig. 2



(a) ground-truth (b) FBP: 21.19, 0.22 (c) TV: 29.85, 0.79 (d) U-net: 34.42, 0.90

(e) LPD: 35.76, 0.92 (f) AR: 33.52, 0.86 (g) ACR: 31.24, 0.86 (h) ACR-SC: 30.93, 0.85

(i) ground-truth (j) FBP: 21.59, 0.24 (k) TV: 29.16, 0.77 (l) U-net: 32.69, 0.87

(m) LPD: 34.05, 0.89 (n) AR: 32.14, 0.84 (o) ACR: 30.14, 0.83 (p) ACR-SC: 29.88, 0.82

Figure 2: CT reconstruction on Mayo clinic data with the respective PSNR (dB) and SSIM scores indicated below. Enforcing the source condition leads to minor deterioration in the performance of the ACR, but we gain in terms of theoretical convergence properties.

indicate that ACR-SC is only marginally inferior to ACR in terms of the quality metrics, while it still outperforms classical hand-crafted regularizers such as total-variation (TV). The slight deterioration as compared to ACR can be attributed to the additional regularity constraint imposed by the SC.

4 Conclusions

We addressed the problem of learning data-adaptive convex regularizers for inverse problems and proposed an approach to augment the existing adversarial learning framework with a suitable penalty that enforces the variational source condition. Incorporating the source condition penalty leads to only minor degradation in the numerical performance, but it simultaneously offers a theoretical grounding for Bregman iterations and paves the way for deriving convergence rate estimates for the resulting variational reconstruction problem as the noise-level approaches zero.

References

[1] J. Adler, H. Kohr, and O. Öktem. Operator discretization library (odl). *Software available from https://github.com/odlgroup/odl*, 2017.

[2] B. Amos, L. Xu, and J. Z. Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155, 2017.

[3] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[4] M. Benning and M. Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, 2018.

[5] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Problems*, pages 1411–1421, 2004.

[6] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028v3*, Dec. 2017.

[7] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[8] E. Kobler, A. Effland, K. Kunisch, and T. Pock. Total deep variation for linear inverse problems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7549–7558, 2020.

[9] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier. NETT: solving inverse problems with deep neural networks. *Inverse Problems*, 36(6), 2020.

[10] S. Lunz, O. Öktem, and C.-B. Schönlieb. Adversarial regularizers in inverse problems. In *Advances in Neural Information Processing Systems*, pages 8507–8516, 2018.

[11] C. McCollough. Tfg-207a-04: Overview of the low dose ct grand challenge. *Medical Physics*, 43(6):3759–3760, 2014.

[12] S. Mukherjee, S. Dittmer, Z. Shumaylov, S. Lunz, O. Öktem, and C.-B. Schönlieb. Learned convex regularizers for inverse problems. *arXiv preprint arXiv:2008.02839v2*, 2021.

[13] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. S. Kazimierski. *Regularization methods in Banach spaces*. de Gruyter, 2012.