Learning to solve Inverse problems

July 9, 2018

Outline

Outline

Inverse Problems

Forward model $\mathbf{Y} = \Phi_\theta(\mathbf{X}) + \textit{Noise}$

Inverse problem Given $\mathbf{Y}$ estimate either $\mathbf{X}$ or parameters denoted by θ in system

Examples

- MRI, CT and other medical imaging modes. Given $\mathbf{Y}$ Fourier measurements and $\mathbf{X}$ illumination system, we estimate θ in person
- Recommender systems. Given $\mathbf{Y}$ the user preferences and θ the factor models we predict $\mathbf{X}$ the rankings on novel objects.

Inverse Problems

Forward model $\mathbf{Y} = \Phi_\theta(\mathbf{X}) + \textit{Noise}$

Inverse problem Given $\mathbf{Y}$ estimate either $\mathbf{X}$ or parameters denoted by θ in system

Examples

- MRI, CT and other medical imaging modes. Given $\mathbf{Y}$ Fourier measurements and $\mathbf{X}$ illumination system, we estimate θ in person
- Recommender systems. Given $\mathbf{Y}$ the user preferences and θ the factor models we predict $\mathbf{X}$ the rankings on novel objects.

Φ is linear in most cases. Ill-posed due to insufficient measurements compared to signal dimensions.

Inverse Problems

Forward model $\mathbf{Y} = \Phi_\theta(\mathbf{X}) + \mathit{Noise}$
Inverse problem Given $\mathbf{Y}$ estimate either $\mathbf{X}$ or parameters denoted by θ in system

Examples

- MRI, CT and other medical imaging modes. Given $\mathbf{Y}$ Fourier measurements and $\mathbf{X}$ illumination system, we estimate θ in person
- Recommender systems. Given $\mathbf{Y}$ the user preferences and θ the factor models we predict $\mathbf{X}$ the rankings on novel objects.

Φ is linear in most cases. Ill-posed due to insufficient measurements compared to signal dimensions.

Solution 101 - invert the known operator using the ML estimation depending on noise distribution

Inverse Problems

Forward model $\mathbf{Y} = \Phi_\theta(\mathbf{X}) + \textit{Noise}$
Inverse problem Given $\mathbf{Y}$ estimate either $\mathbf{X}$ or parameters denoted by θ in system

Examples

- MRI, CT and other medical imaging modes. Given $\mathbf{Y}$ Fourier measurements and $\mathbf{X}$ illumination system, we estimate θ in person
- Recommender systems. Given $\mathbf{Y}$ the user preferences and θ the factor models we predict $\mathbf{X}$ the rankings on novel objects.

Φ is linear in most cases. Ill-posed due to insufficient measurements compared to signal dimensions.
Solution 101 - invert the known operator using the ML estimation depending on noise distribution
Better Solution Impose known conditions on the solution space to restrict the solution space to Ω.

$$\min_{\mathbf{X}\in\Omega} \textit{Loss}\,(X, Y; \Phi) \textbf{ or } \min_{\theta\in\Omega} \textit{Loss}\,(\theta, Y; X)\,.$$

Solution

$$\min_{\mathbf{X}\in\Omega} Loss\left(X, Y; \Phi\right).$$

Some common assumptions are sparsity, low-rank, smoothness, band-limited. We focus on the sparsity condition in this discussion. Under the Gaussian noise assumption, we use the least squares loss with $\mathbf{X} \in \mathcal{D}_K$, where $\mathcal{D}_K$ is the set of all K-sparse vectors in N dimensional space, $\Phi \in \mathbb{C}^{M\times N}$ and $\mathbf{Y} \in \mathbb{C}^M$

$$\min_{\mathbf{X}\in\mathcal{D}_K} \left\|\mathbf{Y} - \Phi\mathbf{X}\right\|^2$$

Cardinality constraint

$$\min_{\mathbf{X}\in\mathcal{D}_K} \left\|\mathbf{Y} - \Phi\mathbf{X}\right\|^2$$

subject to $\|\mathbf{x}\|_0 \leq K$, $\|\mathbf{x}\|_\infty \leq 1.$

Solution

$$\min_{\mathbf{X}\in\Omega} Loss\,(X, Y; \Phi)\,.$$

Some common assumptions are sparsity, low-rank, smoothness, band-limited. We focus on the sparsity condition in this discussion. Under the Gaussian noise assumption, we use the least squares loss with $\mathbf{X} \in \mathcal{D}_K$, where $\mathcal{D}_K$ is the set of all K-sparse vectors in N dimensional space, $\Phi \in \mathbb{C}^{M\times N}$ and $\mathbf{Y} \in \mathbb{C}^M$

$$\min_{\mathbf{X}\in\mathcal{D}_K} \|\mathbf{Y} - \Phi\mathbf{X}\|^2$$

Cardinality constraint

$$\min_{\mathbf{X}\in\mathcal{D}_K} \|\mathbf{Y} - \Phi\mathbf{X}\|^2$$

subject to $\|\mathbf{x}\|_0 \leq K$, $\|\mathbf{x}\|_\infty \leq 1.$

Convex Relaxation

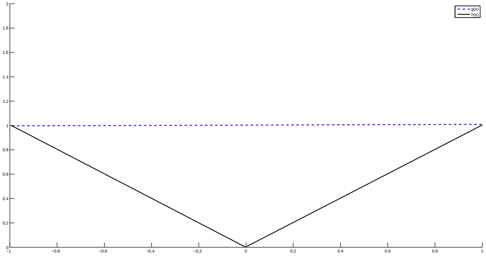
$$\min_{\mathbf{X}\in\mathcal{D}_K} \|\mathbf{Y} - \Phi\mathbf{X}\|^2$$

subject to $\|\mathbf{x}\|_1 \leq K$, $\|\mathbf{x}\|_\infty \leq 1.$

Tightest convex relaxation using Fenchel duality

The definition of the tightest convex relaxation $h(x)$ of any function $g(x), \forall x \in \Omega$ such that $h(x) \leq g(x)$ and $h(x)$ is convex. Fenchel dual of a function is given by $g^*(y) = \sup_x \langle x, y \rangle - g(x)$

Let $g(x) = x^0$ and $h(x) = |x|$ such that $|x| \leq 1$. It can be shown that It can be verified that $g^*(y) = h^*(y)$. Therefore, $g^{**}(z) = h(z)$. This implies that the function $h(.)$ is the tightest convex function.



Technical requirement - Restricted isometry

We define the restricted isometry property of a measurement operator Φ, for all $\mathbf{X} \in \mathcal{D}_K$

$$(1 - \delta_s) \|\mathbf{X}\|^2 \leq \|\mathbf{\Phi X}\|^2 \leq (1 + \delta_s) \|\mathbf{X}\|^2 .$$

In case of the convex relaxation approach the null-space of the measurement operator is important. A bound on δ_{2K} is required to ensure the K sparse solution to be unique [Can08][1].
In case of the greedy methods for imposing the cardinality constraint, the matrix $\Phi^*\Phi$ should be close to an identity matrix. Therefore, a bound on δ_{3K} is sufficient to bound the residual error[BD08][2].

[1] Emmanuel Candes, *The restricted isometry property and its implications for compressed sensing*, Comptes Rendus Mathematique, 2008

[2] Thomas Blumensath, and Mike E. Davies , *Iterative Thresholding for Sparse Approximations*, Journal of Fourier Analysis and Applications,2008.

Outline

Approaches

- Unfolding the iterations of projected gradient methods [XWG^{+}16][3][GL10][4]
- Auto-encoders [MPB15][5]

[3]Bo Xin, Yizhou Wang, Wen Gao, David Wipf, and Baoyuan Wang, *Maximal Sparsity with Deep Networks?*, Advances in Neural Information Processing Systems 29 (NIPS 2016)

[4]Karol Gregor, and Yann LeCun, *Learning fast approximations of sparse coding*, ICML'10 Proceedings of the 27th International Conference on International Conference on Machine Learning

[5]Ali Mousavi , Ankit B. Patel ,and Richard G. Baraniuk, *A deep learning approach to structured signal recovery* 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)

Solving optimization problems with constraints

Unconstrained problem Goal : $\min_x f(x)$
Solution : Each step makes a quadratic approximation and minimize this surrogate

$$x^{t+1} = \arg\min_x f(x^t) + \langle \nabla_x f(x^t), x - x^t \rangle + \frac{1}{2\mu} \left\| x - x^t \right\|^2$$

$$x^{t+1} = x^t - \mu \nabla_x f(x^t)$$

Constrained problem

$$\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x}) = \min_x f(\mathbf{x}) + I_c(x), \qquad (1)$$

$$\text{where } I_{\mathcal{C}}(x) = \begin{cases} 0 & x \in \mathcal{C}, \\ \infty & \textbf{otherwise} \end{cases} .$$

$$x^{t+1} = \Pi_{\mathcal{C}} \left(x^t - \mu \nabla_x f(x^t) \right),$$

$$\Pi_{\mathcal{C}}(z) = \arg\min_{x \in \mathcal{C}} \|x - z\|_2^2 .$$

Projected Gradient descent

Iterative hard thresholding

$$\mathbf{X}^{(t+1)} = \arg\min_{\mathbf{X}} \|\mathbf{Y} - \Phi\mathbf{X}\|^2 + \lambda \|\mathbf{X}\|_0 - \mu \left\|\Phi\left(\mathbf{X} - \mathbf{X}^{(t)}\right)\right\|^2 + \mu \left\|\mathbf{X} - \mathbf{X}^{(t)}\right\|^2$$

$$\mathbf{X}^{(t+1)} = \mathbf{H}_\lambda\left(\left(\mathbf{I} - \mu\Phi^T\Phi\right)\mathbf{X}^{(t)} + \Phi^T Y\right)$$

or

$$\mathbf{X}^{(t+1)} = \mathbf{H}_\lambda\left(\mathbf{X}^{(t)} + \mu\Phi^T Res^{(t)}\right)$$

where the residual is defined as $Res^{(t)} = \mathbf{Y} - \Phi\mathbf{X}^{(t)} = \mathbf{Y} - \Phi\mathbf{X}^{(t-1)} + \Phi\left(\mathbf{X}^{(t)} - \mathbf{X}^{(t-1)}\right)$. Therefore, $Res^{(t)} = Res^{(t-1)} + \Phi\left(\mathbf{X}^{(t)} - \mathbf{X}^{(t-1)}\right)$.

$$\mathbf{H}_\lambda(z) = \begin{cases} 0 & |z| \le \lambda \\ z & |z| > \lambda \end{cases}$$

Convex relaxation

$$\mathbf{X}^{(t+1)} = \arg\min_{\mathbf{X}} \|\mathbf{Y} - \Phi\mathbf{X}\|^2 + \lambda \|\mathbf{X}\|_1 + \mu \left\|\mathbf{X} - \mathbf{X}^{(t)}\right\|^2$$

$$\mathbf{X}^{(t+1)} = \mathbf{S}_\lambda \left(\left(\mathbf{I} - \mu\Phi^T\Phi \right) \mathbf{X}^{(t)} + \Phi^T Y \right)$$

or

$$\mathbf{X}^{(t+1)} = \mathbf{S}_\lambda \left(\mathbf{X}^{(t)} + \mu\Phi^T Res^{(t)} \right)$$

where the residual is defined as $Res^{(t)} = Res^{(t-1)} + \Phi \left(\mathbf{X}^{(t)} - \mathbf{X}^{(t-1)} \right)$.

$$\mathbf{S}_\lambda(z) = \begin{cases} 0 & |z| \leq \lambda \\ \lambda - z & z > \lambda \\ \lambda + z & z > \lambda \end{cases}$$

General steps

$$\mathbf{X}^{(t+1)} = \Psi\left(\mathbf{S}\mathbf{X}^{(t)} + \mathbf{W}Y\right)$$

or

$$\mathbf{X}^{(t+1)} = \Psi\left(\mathbf{X}^{(t)} + \mathbf{W_1}Res^{(t)}\right)$$

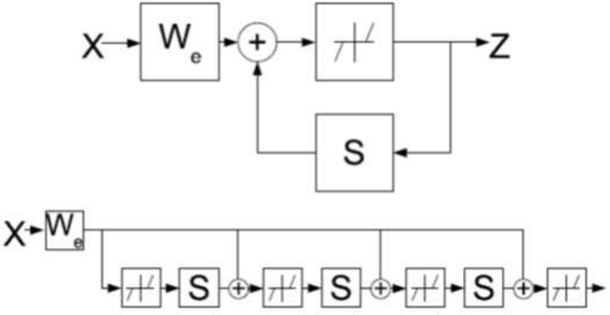
where the residual is defined as $Res^{(t)} = Res^{(t-1)} + \Phi\left(\mathbf{X}^{(t)} - \mathbf{X}^{(t-1)}\right)$ and Ψ is an algorithm specific non-linear function.

Outline

Basic representation of projected gradient step

$$\mathbf{X}^{(t+1)} = \Psi\left(\mathbf{S}\mathbf{X}^{(t)} + \mathbf{W}Y\right)$$

is unrolled over time. The common weights to all the layers are estimated from data.



Overview of the paper

$$\mathbf{X}^{(t+1)} = \Psi\left(\mathbf{S}^{(t)}\mathbf{X}^{(t)} + \mathbf{W}^{(t)}Y\right)$$

1. Provides a preliminary analysis of benefit of adapting weights in improving the RIP constraint
2. Provides a method to learn the inverse map in the setting of correlated dictionary
3. Formulates the sparse recovery problem as a multi-label support recovery problem.

Learning shared weights

$$\mathbf{X}^{(t+1)} = \Psi\left(\mathbf{S}\mathbf{X}^{(t)} + \mathbf{W}Y\right)$$

This update step solves an alternative objective function given by

$$\min_{\mathbf{X}} \frac{1}{2}\mathbf{X}^T\mathbf{W}\Phi\mathbf{X} - \mathbf{X}^T\mathbf{W}\mathbf{Y} \textbf{ subject to } \|\mathbf{X}\| \leq K$$

Let $\mathbf{W} = \mathbf{D}\Phi^{\mathbf{T}}\bar{\mathbf{W}}\bar{\mathbf{W}}^{\mathbf{T}}$. This paper proposes solving the optimization problem indirectly using training samples

$$\min_{\bar{\mathbf{W}},\mathbf{D}} \delta_{3K}\left(\bar{\mathbf{W}}\Phi\mathbf{D}\right) \tag{2}$$

Using training samples and adaptive weights, we can handle certain models of correlation in the measurement matrix by using the weights to get some pre-conditioning.

Learning shared weights - drawbacks

- This projection operator is specific to sparse signal model. This method fails in more structured models such as group sparsity, clustered sparsity, etc.
- Detailed analysis is presented in [GEBS18][6]that generalizes to structured models.

[6] R. Giryes, Y. C. Eldar, A. M. Bronstein and G. Sapiro, "Tradeoffs Between Convergence Speed and Reconstruction Accuracy in Inverse Problems," in IEEE Transactions on Signal Processing, vol. 66, no. 7, pp. 1676-1690, April, 1 2018

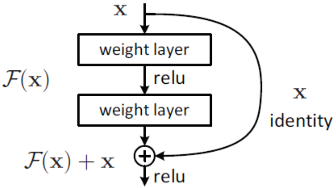
Learning iteration dependent weights

$$\mathbf{X}^{(t+1)} = \Psi\left(\mathbf{S}^{(t)}\mathbf{X}^{(t)} + \mathbf{W}^{(t)}Y\right)$$

Two strategies are used to implement this learning strategy

- Residual networks
- Long-short term memory networks.

Residual networks



Residual neural network learns a function that decomposes as

$$\mathcal{H}(x) = \Psi(\mathcal{F}(x) + x)$$

This step is similar to an iteration in the projected gradient descent method where Ψ is the rectified linear functional.

$$\mathbf{X}^{(t+1)} = \Psi\left(\mathbf{X}^{(t)} + \mathbf{W}_1 Res^{(t)}\right)$$

is the most suited formulation to use with the residual networks. This also enables to use Residual from previous step rather than the input itself as shown

$$Res^{(t)} = Res^{(t-1)} + \Phi\left(\mathbf{X}^{(t)} - \mathbf{X}^{(t-1)}\right)$$

LSTM formulation

- This formulation is discussed in detail in [HXIW17][7].
- This formulation uses the re-weighted least squares formulation to solve the sparsity constraint problem. This is similar to FOCUSS or SBL.
- The correspondence between the gates in the LSTM and the components of the optimization program are established. More of this will be discussed later.

[7] Hao He, Bo Xin, Satoshi Ikehata, and David Wipf, "From Bayesian Sparsity to Gated Recurrent Nets," Advances in Neural Information Processing Systems (NIPS), 2017

References I

Thomas Blumensath and Mike E. Davies, *Iterative thresholding for sparse approximations*, Journal of Fourier Analysis and Applications **14** (2008), no. 5, 629–654.

E. Candes, *The restricted isometry property and its implications for compressed sensing*, Comptes Rendus Mathematique **346** (2008), no. 9-10, 589–592.

R. Giryes, Y. C. Eldar, A. M. Bronstein, and G. Sapiro, *Tradeoffs between convergence speed and reconstruction accuracy in inverse problems*, IEEE Transactions on Signal Processing **66** (2018), no. 7, 1676–1690.

Karol Gregor and Yann LeCun, *Learning fast approximations of sparse coding*, Proceedings of the 27th International Conference on International Conference on Machine Learning (USA), ICML'10, Omnipress, 2010, pp. 399–406.

Hao He, Bo Xin, Satoshi Ikehata, and David Wipf, *From bayesian sparsity to gated recurrent nets*, Advances in Neural Information Processing Systems 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), Curran Associates, Inc., 2017, pp. 5554–5564.

A. Mousavi, A. B. Patel, and R. G. Baraniuk, *A deep learning approach to structured signal recovery*, 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Sept 2015, pp. 1336–1343.

Bo Xin, Yizhou Wang, Wen Gao, David Wipf, and Baoyuan Wang, *Maximal sparsity with deep networks?*, Advances in Neural Information Processing Systems 29 (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), Curran Associates, Inc., 2016, pp. 4340–4348.