Inference with l_0-norm-based Sparsity Prior on Discrete Framework

Kyong Joon Lee[1], Dongjin Kwon[1], Il Dong Yun[2], and Sang Uk Lee[1]
[1]School of EECS, Seoul Nat'l Univ., Seoul, 151-742, Korea
[2]School of EIE, Hankuk Univ. of F. S., Yongin, 449-791, Korea
kjoon@cvl.snu.ac.kr, djk@cvl.snu.ac.kr, yun@hufs.ac.kr, sanguk@ipl.snu.ac.kr

Abstract

We present a new penalizing scheme for a recently introduced prior model [8] on discrete frameworks. The model convincingly assumes that the optimal solutions for the frameworks possess sparse representation *on certain transform domains, and applies this sparsity assumption as a prior information for inference problems. Promoting the sparsity, we proposes to penalize l_0-norm of coefficient vector of the transform bases, instead of l_1-norm employed in that recent work. Experiments compare the proposed prior with previous ones and show enhanced performance, both in qualitative and quantitative manner.*

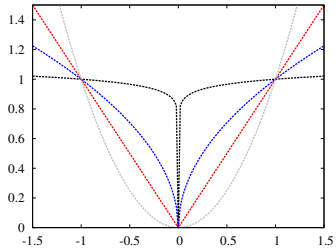


Figure 1: Plots for l_p norms: $p = 2$(gray), $p = 1$(red), $p = 0.5$(blue) and $p = 0.05$(black), respectively. When p approaches to zero, the function becomes severely non-convex.

1. Introduction

Sparse representation of signals has been extensively studied for decades. The main idea is that signals can be represented by linear combination of few (sparse) components in a dictionary containing prototype signals. This well-known property has presented numerous applications in various fields; *e.g.*, lossy compression and compressive sensing [3] are definite examples. In the literature of computer vision, researchers have applied this property to regularization in inverse problems such as denoising [5] and super-resolution [12]. These inverse problems are generally reduced to solve the optimization problem of following form:

$$\min_{\alpha} ||\alpha||_0 \text{ s.t. } \mathbf{x} = \mathbf{\Psi}\alpha, \tag{1}$$

where $\mathbf{x}$ is an observed signal, Ψ is the dictionary matrix. α means a coefficient vector for the linear combination and $||\alpha||_0$ is the l_0-norm of the vector α indicating the number of its non-zero elements.

Exactly solving this equation is reported to be NP-hard, thus a relaxed approach [4] was proposed to address this challenge: By replacing non-convex l_0-norm with convex l_1-norm, the method finds optimal solution using convex optimization [1]. It was also introduced [2] that for many problems, minimizing the l_1-norm is equivalent to minimizing the l_0-norm under certain conditions. However, experimental results in [9] posed a strong question about the equivalence of minimizing both norms in practical problems.

Meanwhile, a recent study [8] proposed a new model applying the sparse constraint for the inverse problems defined on discrete frameworks. This model employs discrete optimization strategy [11] which has been presenting state-of-the-art performance in various applications. In contrast to the previous methods based on convex optimization, it presents a big advantage to address more flexible energy terms; *e.g.*, non-convex (but more robust) cost functions.

Despite that advantage, its energy model still stays on using l_1-norm for the sparsity-promoting term, showing unsatisfactory results in applications. To this end we propose to employ the original term; *i.e.*, l_0-norm, enforcing the exact meaning of sparsity. Our experiments show that this simple strategy greatly enhances the performance without any additive computational complexity.

2. Proposed Prior Model

We consider a problem finding the MAP (Maximum-a-Posteriori) for a discrete random field. Let $\mathcal{G}$ be an undirected graph with node set $\mathcal{V}$ and clique set $\mathcal{C}$. Let x_s be

a random variable in some discrete sample space $\mathcal{X}_s = \{0, \ldots, L-1\}$, representing the label of the node $s \in \mathcal{V}$. Provided the posterior follows the Gibbs distribution, we convert the problem into minimizing energy functional defined with likelihood and prior potentials as follows:

$$E(\mathbf{x}) = \sum_{s \in \mathcal{V}} \theta_s(x_s) + \sum_{c \in \mathcal{C}} \theta_c(\mathbf{x}_c). \tag{2}$$

When the clique set only involves an edge set $\mathcal{E} = \{(s,t)|s,t \in \mathcal{V}\}$ and the potential is defined as $\theta_c(x_s, x_t) = \rho(x_s - x_t)$, the prior is well-known smoothness prior with pairwise potential term. Likewise we can define other prior models in the previous works with higher order clique potentials.

The proposal in [8] starts from a novel prior assumption: the optimal label configuration can be represented by sparse combination of basis signals on a transform domain. Strictly, it is the signal mapped from the label configuration that has the sparse representation, not the configuration itself. However, for notational and conceptual simplicity, we may assume the mapping is linear: $\tau_s(x_s) = ax_s + b_s$ where a is a scale factor and b_s is an offset. Then following equations are all valid and only differ by the scale and offset.

We may also assume that this prior knowledge is still applicable even for smaller parts of the solution. A part of the solution $\hat{\mathbf{x}}$ $(= \arg\min_{\mathbf{x}} E(\mathbf{x}))$ is assumed to be represented by combination of few basis label configurations, shown as follows:

$$\hat{\mathbf{x}}_c = \mathbf{\Psi}\alpha_c, \tag{3}$$

where $\mathbf{\Psi}$ is an orthonormal complete matrix whose columns are the basis solutions, and α_c is a coefficient vector.

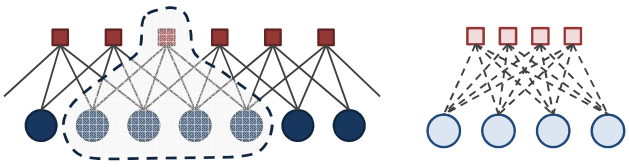
Most of components of α_c should be zero to be referred as *sparse*. We penalize the non-zero terms using l_p norm.

$$\begin{aligned} \theta_c(\mathbf{x}_c) = ||\alpha_c||_p^p &= \sum_{i=1}^{N} |\alpha_c^{(i)}|^p \\ &= \sum_{i=1}^{N} |\psi_i^T \mathbf{x}_c|^p, \end{aligned} \tag{4}$$

where $\alpha_c^{(i)}$ means i^{th} component of α_c, ψ_i^T means the transpose of i^{th} column of Ψ, and N indicates the size of the clique $\mathbf{x}_c$. Figure 2b presents graphical structure of this energy model when $N = 4$.

The penalizer better promotes sparsity when $p \to 0$; but it becomes severely non-convex as shown in Figure 1. Taking advantage of the discrete framework, we propose to use the l_0-norm as it is, less concerning local minima. The final energy formulation is shown as follows:

$$E(\mathbf{x}) = \sum_{s \in \mathcal{V}} \theta_s(x_s) + \sum_{c \in \mathcal{C}} \sum_{i=1}^{N} [\psi_i^T \mathbf{x}_c = 0], \tag{5}$$



(a) (b)

Figure 2: Graphical illustration for the proposed model. (a) Overlapped cliques ($N = 4$) are shown in dark red squares while nodes are shown in dark blue circle. (b) Each clique is sum of cliques (shown in light red squares) with special form; known as linear constraint nodes [10].

where $[\cdot]$ is one if its argument is true and zero otherwise.

3. Optimization

The clique size N needs to be large enough in order for the sparsity works as a prior. Thus the proposed prior potential involves extremely high order potentials. Minimizing that high order potential in general is not feasible under current hardware capabilities. In addition, overlapped cliques in the graph yield very complicated structure for optimization, as shown in 2a.

To this challenge, we employ the method proposed in [8]: we apply the dual decomposition [6] for overall optimization. Dual decomposition is an optimization method obtaining the solution from a difficult large problem by decomposing it into smaller subproblems; and then combining the solutions addressing the subproblems. We decompose each of the clique potentials into an individual subproblem. Resolving the high order potentials in the subproblems, we apply the efficient message-passing method using the property that clique potentials are linear constraint [10]. Remarkably, these procedures can be implemented on parallel hardware providing a practical framework.

4. Experiments

Experiments are designed to show performance enhancement by employing l_0-norm rather than l_1; although we have also provided results from several smoothness priors. For algorithmic simplicity [8], we use the WHT (Walsh-Hadamard Transform) as basis matrix.

4.1. Signal reconstruction

We start with conducting simple 1-D signal reconstruction tests. We consider 5 different types of groundtruth signals; *i.e.*, Step4, Step2, Slope, Sawtooth and Half-Circle. We generate 10 noisy input signals for each, to see average performance. The input signals are generated by adding Gaussian noise with $\sigma = 8$. The amount of noise is cropped in $[-5, 5]$ to limit the number of label $L = 11$. For unary

Table 1: PSNR results for signal reconstruction ($\sigma = 8$.) (1) Step4, (2) Step2, (3) Slope, (4) Sawtooth and (5) Half-Circle.

Prior	PSNR (dB)				
	(1)	(2)	(3)	(4)	(5)
1^{st}-Order	37.34	35.12	36.74	37.76	37.43
1^{st}+2^{nd}+3^{rd}	37.73	44.83	39.72	38.69	41.09
Sparsity(l_1)	53.29	49.77	37.81	36.48	34.21
Sparsity(l_0)	59.74	56.29	39.10	39.21	40.13

potentials, we use $\theta_s(x_s) = |y_s - \tau_s(x_s)|$ where y_s is a value at s of the input signal and $\tau_s(x_s) = x_s + b_s$ where $x_s \in \{0, \ldots, 10\}$. We set $b_s = y_s - 5$ which means we find the solution in the range of $[y_s-5, y_s+5]$. The clique size N is set to 16 while total length of input signal is 64. Resulting graph contains 49 subtrees where each subtree contains 16 linear constraint cliques. More detailed configurations are referred to the previous work [8].

Figure 3 shows qualitative results. Upper two rows present the cases that contain only step functions. The 1^{st} order prior (b) generates locally regularized effect while the combination prior (c) generally performs better but over-fits curves around edges. In contrast, both of the sparsity priors (d,e) yield outperforming results on homogeneous as well as edge region; however we may conclude the proposed prior gives more robust performance, as we can find some extra perturbations in (d). This can be clearly seen in the quantitative results (PSNR) in Table 1.

For the rest of cases where the sparsity priors are not expected to show the best performance (due to inherent limitation of the WHT), we still found the proposed prior yields much better results than l_1-norm-based one; moreover, it performs almost close to (Line, Half-Circle) or better than (Sawtooth) the combination prior.

4.2. Image denoising

We extend the reconstruction problem to 2-D domain. We use synthetic 32×32 images with plain structure (Figure 4.) Detail configuration such as additive noise, number of labels, unary potential just follows that of Section 4.1.

Since images are two dimensional, two different type of cliques can be overlapped on the nodes; *i.e.*, 1-D line and 2-D patch types. For simplicity we only tested the line-type. Solving the line-type clique can be considered as solving multiple signal reconstruction in parallel along with x and y coordinates. We set clique size $N = 16$ producing 1088 cliques.

As shown in Figure 4 and average PSNR, the proposed prior presents the best performance, the combination prior is next and the l_1-norm-based one follows. Considering the line-type is direct extension of 1-D signal construction and the image is also the case of 1-D heavyside step function,

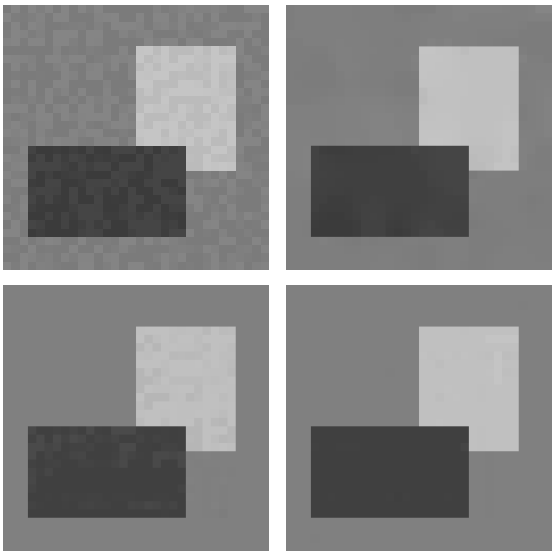


Figure 4: **Top-left:** Noisy 32×32 input image (PSNR: 36.21). Denoising result from **Top-right:** combination of 1^{st}, 2^{nd} and 3^{rd} order smoothness prior (PSNR: 41.11), **Bottom-left:** l_1-norm-based prior (PSNR: 40.42), and **Bottom-right:** the proposed prior (PSNR: 54.64). Best viewed electronically.

it is notable that the l_1-norm based prior yields worse result than the combination prior. (Since it produced much better results in 1-D case.)

This result implies that when the graph structure becomes highly complex, the robustness of sparsity-promoting term much influences the performance; and thus the proposed l_0-norm prior is definitely preferable in real applications. Both sparsity priors take the same 10.1 seconds per iteration, which is tolerable time considering the image and clique size.

5. Conclusion

This paper proposed a new sparsity-promoting penalizer for the discrete framework. Our work much improved the performance of the previous work by simply employing the l_0-norm instead of the l_1-norm. This result indicates that the equivalence of minimizing both norms [2] may not hold true in real applications as was already issued in [9]. We plan to further apply the proposed method to various problems to convince the difference.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 1

[2] E. J. Candés and T. Tao. Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies?

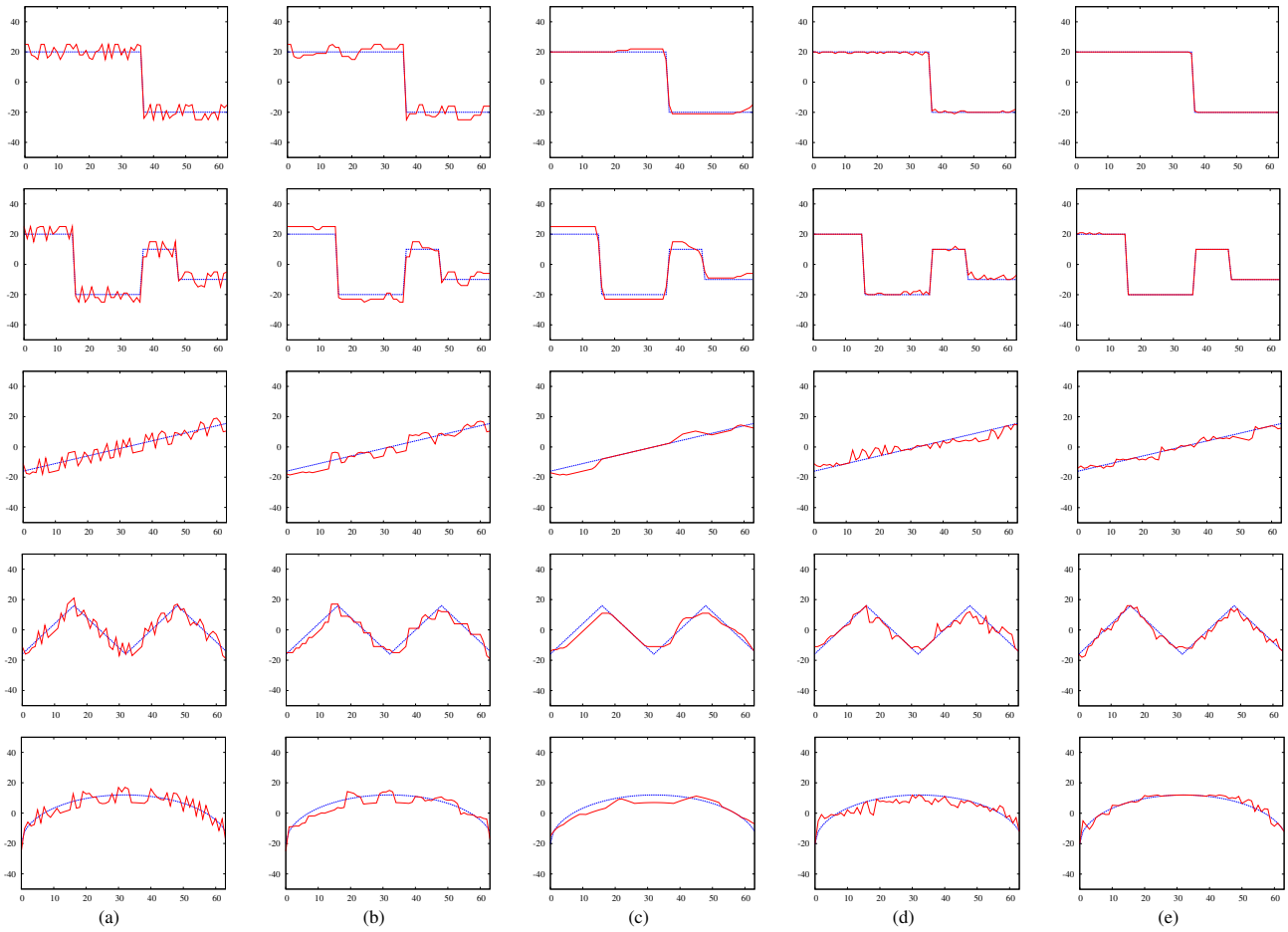


Figure 3: Signal reconstruction results using different prior information. (a) Noisy input signal with blue dotted line illustrating groundtruth signal. Reconstruction results using (b) 1^{st} order smoothness prior, (c) combination of 1^{st}, 2^{nd} and 3^{rd} order smoothness prior [7], (d) l_1-norm-based prior [8], and (e) the proposed prior.

IEEE Transactions on Information Theory, 52(12):5406–5425, 2006. 1, 3

[3] E. J. Candés and M. B. Wakin. Introduction to Compressive Sampling. *IEEE Signal Processing Magazine*, 21, 2008. 1

[4] D. L. Donoho. For Most Large Undetermined Systems of Linear Equations the Minimal l_1-norm Solution is also the Sparsest Solution. Technical Report http://www-stat.stanford.edu/donoho/Reports/, Sep. 2004. 1

[5] M. Elad and M. Aharon. Image Denoising via Sparse and Redundant Representations over Learned Dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006. 1

[6] N. Komodakis, N. Paragios, and G. Tziritas. MRF Energy Minimization and Beyond via Dual Decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:531–552, 2011. 2

[7] D. Kwon, K. J. Lee, I. D. Yun, and S. U. Lee. Solving MRFs with Higher-Order Smoothness Priors Using Hierarchical Gradient Nodes. In *Proc. ACCV*, 2010. 4

[8] K. J. Lee, D. Kwon, I. D. Yun, and S. U. Lee. Sparse Random Field: Discrete Random Field Model with Sparsity Prior. Technical Report http://spl.snu.ac.kr/spl/publications/reports/, Seoul National Univ., Dept. of EECS, Feb. 2011. 1, 2, 3, 4

[9] L. Mancera and J. Portilla. L0-norm-based Sparse Representation through Alternative Projections. In *Proc. ICIP*, 2006. 1, 3

[10] B. Potetz and T. S. Lee. Efficient Belief Propagation for Higher-Order Cliques using Linear Constraint Nodes. *Computer Vision and Image Understanding*, 112(1):39–54, 2008. 2

[11] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A Comparative Study of Energy Minimization Methods for Markov Random Fields. In *Proc. 9th ECCV*, 2006. 1

[12] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2008. 1