Stability Conditions for L_1/L_p Regularization

Tom Heskes and Marcel van Gerven

1 Introduction

L_1/L_p regularization is a regularization approach that has the same sparsifying properties as L_1 regularization and allows regularization over feature *groups* instead of features [1]. This approach is of use when features can be partitioned into groups that are seen as belonging together as well as in the case of transfer learning, where the same features are grouped together over multiple tasks. In this note, we derive stability conditions that determine for which value of a regularization parameter λ a group of feature will be included into the model.

2 L_1/L_p regularization

In L_1/L_p regularization, we consider cost functions of the form

$$E(\Theta) = L(\Theta) + \lambda \|\Theta\|_{1,p} \,,$$

with Θ a matrix with components θ_{ij}, $L(\Theta)$ a loss term, such as the negative loglikelihood of data given a particular probability model parameterized by Θ, λ a regularization constant, and the L_1/L_p norm

$$||\Theta||_{1,p} \equiv \sum_{i=1}^{I} \sqrt[p]{\sum_{j=1}^{J} |\theta_{ij}|^p} \,.$$

We write $\boldsymbol{\theta}_i$ for the vector with components θ_{ij}. The L_1/L_p norm uses an L_1 norm to regularize over vectors $\boldsymbol{\theta}_i$ and an L_p norm to regularize *within* vectors $\boldsymbol{\theta}_i$. If we choose $p = 1$ then we obtain

$$||\Theta||_{1,p} = \sum_{i=1}^{I} \sum_{j=1}^{J} |\theta_{ij}|$$

which amounts to standard L_1 regularization over components θ_{ij}. It is well-known that by penalizing differences from zero, L_1 regularization effectively performs feature selection with respect to components θ_{ij}. However, if $p > 1$ then the components θ_{ij} within a vector $\boldsymbol{\theta}_i$ become tied, leading to feature selection with respect to vectors $\boldsymbol{\theta}_i$. In the limit, as p goes to infinity, the regularization of a vector $\boldsymbol{\theta}_i$ is fully determined by the component θ_{ij} having the largest magnitude. L_1/L_p regularization has the advantage over plain L_1 regularization that we may group together components that are seen as belonging together. For example, if $L(\Theta)$ is interpreted as minus the loglikelihood of data given a particular probability model parameterized by Θ, i runs over features, and j over tasks, then we may enforce that the same features are used for solving different tasks (this is known as transfer learning or multi-task learning).

When we regard the vectors $\boldsymbol{\theta}_i$ as features then it becomes useful to determine for which values of λ the solution $\boldsymbol{\theta}_i = \mathbf{0}$ remains stable. Consider fixing all vectors $\boldsymbol{\theta}_l$ for $l \neq i$ to some $\boldsymbol{\theta}_l^*$ (which may or may not be equal to the null vector) and possibly changing $\boldsymbol{\theta}_i$ away from $\mathbf{0}$. We define

$$\Theta_i^*(\boldsymbol{\theta}) \equiv [\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{i-1}^*, \boldsymbol{\theta}, \boldsymbol{\theta}_{i+1}^*, \ldots, \boldsymbol{\theta}_I^*] \,.$$

It can then be shown that the solution $\Theta_i^*(\mathbf{0})$ is stable if and only if

$$\lambda \geq \left\| \left. \frac{\partial L(\Theta)}{\partial \boldsymbol{\theta}_i} \right|_{\Theta=\Theta_i^*(\mathbf{0})} \right\|_{p/(p-1)} . \tag{1}$$

Loosely speaking, to move the parameters away from zero the push due to the gradient of the loss term should be stronger than the pull of the regularization term.

Proof. Since the vectors $\boldsymbol{\theta}_l$ are fixed for $l \neq i$, we can restrict ourselves to study the dependency of $E(\Theta)$ on $\boldsymbol{\theta}_i$:

$$E(\boldsymbol{\theta}_i) \equiv E(\Theta_i^*(\boldsymbol{\theta}_i)) = L(\Theta_i^*(\boldsymbol{\theta}_i)) + \lambda \|\boldsymbol{\theta}_i\|_p + \text{ constant} .$$

The solution $\boldsymbol{\theta}_i = \mathbf{0}$ is stable if it holds that

$$E(\boldsymbol{\theta}_i) \geq E(\mathbf{0}) \quad \text{for any choice of infinitesimal } \boldsymbol{\theta}_i. \tag{2}$$

A first order Taylor expansion for $\boldsymbol{\theta}_i$ close to $\mathbf{0}$ yields:

$$E(\boldsymbol{\theta}_i) = E(\mathbf{0}) + \sum_{j=1}^{J} \theta_{ij} \left. \frac{\partial L(\Theta)}{\partial \theta_{ij}} \right|_{\Theta=\Theta_i^*(\mathbf{0})} + \lambda \|\boldsymbol{\theta}_i\|_p \equiv E(\mathbf{0}) + \boldsymbol{g}^T \boldsymbol{\theta}_i + \lambda \|\boldsymbol{\theta}_i\|_p ,$$

where here and in the following we ignore higher order terms and we defined, for ease of notation, $\boldsymbol{g} \equiv \left. \frac{\partial L(\Theta)}{\partial \boldsymbol{\theta}_i} \right|_{\Theta=\Theta_i^*(\mathbf{0})}$. The condition (2) thus boils down to

$$\lambda \geq -\frac{\boldsymbol{g}^T \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_p} \quad \forall_{\boldsymbol{\theta}} ,$$

and thus to

$$\lambda \geq \max_{\boldsymbol{\theta}} \left[-\frac{\boldsymbol{g}^T \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_p} \right] = \max_{\boldsymbol{\theta}} \frac{\boldsymbol{g}^T \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_p} = \max_{\boldsymbol{\theta}; \|\boldsymbol{\theta}\|_p = 1} \boldsymbol{g}^T \boldsymbol{\theta} . \tag{3}$$

The last step follows from the observation that the function to be minimized is insensitive to scaling of $\boldsymbol{\theta}$ (as it should) and we can therefore constrain the norm of $\boldsymbol{\theta}$ to any arbitrary value (here chosen to be 1). The first follows from a symmetry argument. Introducing a Lagrange multiplier γ for this constraint, we get the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \gamma) = \boldsymbol{g}^T \boldsymbol{\theta} + \gamma(\|\boldsymbol{\theta}\|_p - 1) ,$$

with derivative

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \gamma)}{\partial \theta_j} = g_j + \gamma \|\boldsymbol{\theta}\|_p^{1/p-1} |\theta_j|^{p-1} \operatorname{sgn}(\theta_j) .$$

Setting this derivative to zero, we see that the optimal solution $\boldsymbol{\theta}^*$ obeys (for $p > 1$)

$$\theta_j^* \propto |g_j|^{1/(p-1)} \operatorname{sgn}(g_j) .$$

Furthermore, we see that we obtain a maximum when the proportionality constant is positive, and a minimum when it is negative. Plugging this into (3), we finally obtain, after some rewriting

$$\lambda \geq \|\boldsymbol{g}\|_{p/(p-1)} .$$

□

Note that if we choose $p = 1$ in (1) then we obtain

$$\lambda \geq \left\| \left. \frac{\partial L(\Theta)}{\partial \boldsymbol{\theta}_i} \right|_{\Theta=\Theta_i^*(\mathbf{0})} \right\|_{\infty} = \max_i \left| \left. \frac{\partial L(\Theta)}{\partial \boldsymbol{\theta}_i} \right|_{\Theta=\Theta_i^*(\mathbf{0})} \right| ,$$

which is the stability condition for L_1 regularization [2].

References

[1] G. Obozinski, B. Taskar, and M. I. Jordan. Multi-task feature selection. Technical report, UC Berkeley, Berkeley, 2006.

[2] S. Perkins, K. Lacker, and J. Theiler. Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research*, 3:1333–1356, 2003.