A Note on k-support Norm Regularized Risk Minimization

Matthew Blaschko

A Note on k-support Norm Regularized Risk Minimization

Matthew B. Blaschko

École Centrale Paris
Grande Voie des Vignes
92295 Châtenay-Malabry, France
matthew.blaschko@inria.fr

Abstract. The k-support norm has been recently introduced to perform correlated sparsity regularization [1]. Although Argyriou et al. only reported experiments using squared loss, here we apply it to several other commonly used settings resulting in novel machine learning algorithms with interesting and familiar limit cases. Source code for the algorithms described here is available from https://github.com/blaschko/ksupport.

Keywords: k-support norm, structured sparsity, regularization, least-squares, hinge loss, support vector machine, SVM, regularized logistic regression, AdaBoost, support vector regression, SVR

1 The k-support Norm

The k-support norm is the gauge function associated with the convex set

$$\operatorname{conv}\{\beta \mid \|\beta\|_0 \leq k, \|\beta\|_2 \leq 1\}. \tag{1}$$

It can be computed as

$$\|\beta\|_k^{sp} = \left(\sum_{i=1}^{k-r-1} (|\beta|_i^{\downarrow})^2 + \frac{1}{r+1} \left(\sum_{i=k-r}^{d} |\beta|_i^{\downarrow} \right)^2 \right)^{\frac{1}{2}} \tag{2}$$

where $|\beta|_i^{\downarrow}$ is the ith largest element of the vector and r is the unique integer in $\{0, \dots, k-1\}$ satisfying

$$|\beta|_{k-r-1}^{\downarrow} > \frac{1}{r+1} \sum_{i=k-r}^{d} |\beta|_i^{\downarrow} \geq |\beta|_{k-r}^{\downarrow}. \tag{3}$$

We use the following notation here: $X \in \mathbb{R}^{n \times d}$ is a design matrix of n samples each with d dimensions; $y \in \mathbb{R}^n$ is the vector of targets.

In the case that $k = 1$ the k-support norm is exactly equivalent to the ℓ_1 norm. In the case that $k = d$, where $\beta \in \mathbb{R}^d$, the k-support norm is equivalent to the ℓ_2 norm.

We note that for an objective

$$\min_{\beta} \lambda \|\beta\|_k^{sp} + f(\beta, X, y) \tag{4}$$

with some loss function $f(\cdot, \cdot, \cdot)$, when $k = d$, this is equivalent to

$$\min_{\beta} \lambda \|\beta\|_2 + f(\beta, X, y) \tag{5}$$

rather than the familiar squared ℓ_2 regularizer. However, for any λ there exists some $\tilde{\lambda}$ such that

$$\arg\min_{\beta} \lambda \|\beta\|_2 + f(\beta, X, y) = \arg\min_{\beta} \tilde{\lambda} \|\beta\|_2^2 + f(\beta, X, y). \tag{6}$$

This can be easily seen by noting that the objectives are the Lagrangians of constrained minimization problems that minimize f subject to the equivalent constraints $\|\beta\|_2 \leq B$ and $\|\beta\|_2^2 \leq B^2$, respectively, for some $B \in \mathbb{R}_+$.

2 Squared Loss

If we use Nesterov's accelerated method (a first-order proximal algorithm) for optimization as suggested in [1], a given implementation of k-support regularized risk requires a function that computes the loss f, a function that computes the gradient of the loss function $\frac{\partial f}{\partial \beta}$, and the Lipschitz constant L for $\frac{\partial f}{\partial \beta}$. We assume that f is convex and differentiable everywhere and that L is finite.

For the squared loss:

$$\begin{aligned} f_2(\beta, X, y) &= \|X\beta - y\|^2 && (7) \\ \frac{\partial f_2}{\partial \beta} &= 2X^T X\beta - 2X^T y && (8) \\ L_2 &= 2\gamma && (9) \end{aligned}$$

where γ is the largest eigenvalue of $X^T X$.

The objective function

$$\lambda \|\beta\|_k^{sp} + \|X\beta - y\|^2 \tag{10}$$

clearly has the lasso [11] and ridge regression [12] as special cases when $k = 1$ and $k = d$, respectively. Argyriou et al. [1] have previously discussed the relationship to the elastic net [17]. The k-support norm with squared loss has been shown to give good results on fMRI data [7].

3 One Sided Squared Loss

While we have previously assumed that $y \in \mathbb{R}^n$, here we will assume we are dealing with the binary classification case where $y \in \{-1, +1\}^n$. One sided squared loss simply computes the squared loss when a margin is violated, and zero otherwise.

$$
\begin{align}
f_{2-}(\beta, X, y) &= \sum_{i=1}^{n} (\max\{1 - y_i \langle \beta, x_i \rangle, 0\})^2 \tag{11}\\
\frac{\partial f_{2-}}{\partial \beta} &= \sum_{i=1}^{n} \begin{cases} 0 & \text{if } y_i\langle \beta, x_i\rangle > 1 \\ 2\langle \beta, x_i \rangle x_i - 2 y_i x_i & \text{if } y_i \langle \beta, x_i \rangle \leq 1 \end{cases} \tag{12}\\
L_{2-} &= 2\gamma. \tag{13}
\end{align}
$$

One sided squared loss has been considered, for example, in [4].

4 Hinge Loss

Hinge loss is not differentiable, so we apply a Huber approximation to hinge loss [4].[1] The Huber parameter is denoted h:

$$
\begin{align}
f_h(\beta, X, y) &= \sum_{i=1}^{n} \begin{cases} 0 & \text{if } y_i\langle \beta, x_i\rangle > 1 + h \\ \frac{(1+h-y_i\langle \beta, x_i\rangle)^2}{4h} & \text{if } |1 - y_i\langle \beta, x_i\rangle| \leq h \\ 1 - y_i\langle \beta, x_i \rangle & \text{if } y_i\langle \beta, x_i\rangle < 1 - h \end{cases} \tag{14}\\
\frac{\partial f_h}{\partial \beta} &= \sum_{i=1}^{n} \begin{cases} 0 & \text{if } y_i\langle \beta, x_i\rangle > 1 + h \\ \frac{\langle \beta, x_i\rangle x_i - (1+h) y_i x_i}{2h} & \text{if } |1 - y_i\langle \beta, x_i\rangle| \leq h \\ -y_i x_i & \text{if } y_i\langle \beta, x_i\rangle < 1 - h \end{cases} \tag{15}\\
L_2 &= \frac{\gamma}{2h} \tag{16}
\end{align}
$$

where γ is as before the largest eigenvalue of $X^T X$. We note that the Lipschitz constant is in a sense conservative in that it grows with the inverse of h, while we might expect a smaller fraction of the data to actually fall within the quadratic portion of the data. Nevertheless for h not too small, we have not observed any convergence issues with Nesterov's accelerated method. While a small value of h may be desirable in a kernelized setting, here we desire Hinge loss not for sparsity of a dual coefficient vector (indeed the k-support norm does not admit a representer theorem [2]), but rather that the loss not grow more than linearly while remaining convex. In other words, we use the hinge loss primarily for its increased robustness over other losses such as (one-sided) squared loss.

The limit cases are the support vector machine (SVM) [5] when $k = d$ and the ℓ_1 regularized SVM [16] when $k = 1$. The k-support regularized SVM can be seen as an alternative to the elastic net regularized SVM [14], but with a tighter convex relaxation to correlated sparsity (Equation (1)).

[1] Although it is perhaps more natural to incorporate non-differentiable losses with the k-support regularizer in a proximal splitting approach, we have arbitrarily closely approximated non-differentiable losses by differentiable ones for the sake of uniformity of presentation and software implementation.

5 Logistic Loss

Logistic loss is derived from logistic regression, and its minimization is equivalent to logistic regression in the case that it is unregularized [8].

$$f_{\log}(\beta, X, y) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right) \tag{17}$$

$$\frac{\partial f_{\log}}{\partial \beta} = -\sum_{i=1}^{n} \frac{e^{-y_i \langle \beta, x_i \rangle}}{1 + e^{-y_i \langle \beta, x_i \rangle}} y_i x_i \tag{18}$$

$$L_{\log} = \frac{\gamma}{4} \tag{19}$$

where the Lipschitz constant has a factor $\frac{1}{4}$ from the Lipschitz constant of the sigmoid in $\frac{\partial f_{\log}}{\partial \beta}$. k-support regularized regression specializes to previously used regularized logistic regression objectives [9] when $k = 1$ or $k = d$.

6 Exponential Loss

Exponential loss is known primarily through its use in AdaBoost.M1 [6, 8].

$$f_{\exp}(\beta, X, y) = \sum_{i=1}^{n} e^{-y_i \langle \beta, x_i \rangle} \tag{20}$$

$$\frac{\partial f_{\exp}}{\partial \beta} = -\sum_{i=1}^{n} e^{-y_i \langle \beta, x_i \rangle} y_i x_i \tag{21}$$

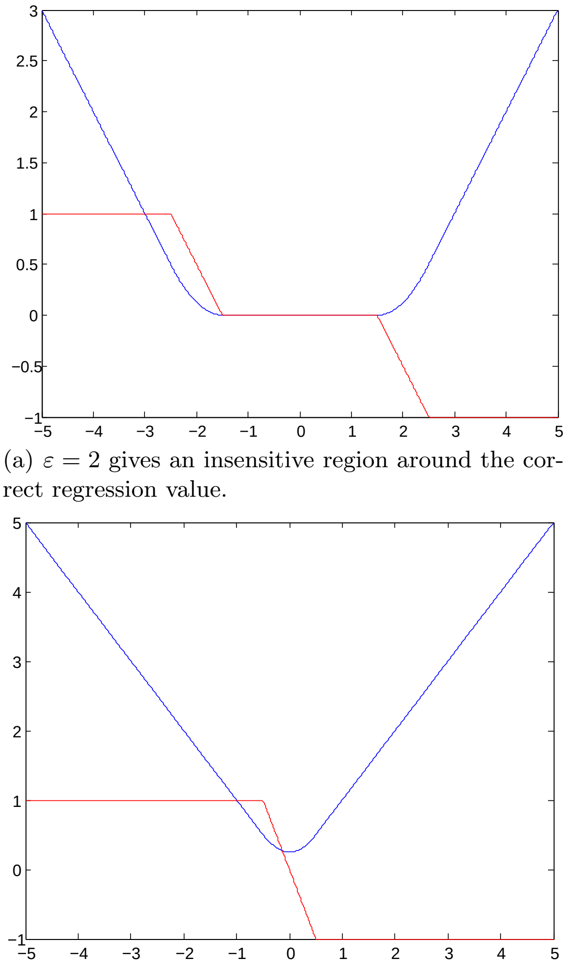
Here, the loss is not globally Lipschitz continuous. However, one may attempt to estimate a sufficiently large constant if one were to apply learning with the k-support norm and Nesterov's accelerated method (we have simply used a relatively conservative $50 \times \gamma$ in the experiments reported in Section 8). As exponential loss is highly degenerate in the presence of label noise (essentially for the same reason that it is not globally Lipschitz continuous), this is likely of limited utility in real-world applications. We have included this loss here primarily for completeness, and have not explored any other optimization strategies.

7 ε-insensitive Loss and Huber Smoothed Absolute Loss

ε-insensitive loss is defined to be [13]:

$$|y_i - \langle \beta, x_i \rangle|_{\varepsilon} := \max\{0, |y - \langle \beta, x_i \rangle| - \varepsilon\} \tag{22}$$

for some parameter $\varepsilon \geq 0$. While ε has an important role in the sparsity of the dual representation for support vector regression [13], that role is not required in the primal. As with hinge loss, we use Huber smoothing to guarantee



(a) $\varepsilon = 2$ gives an insensitive region around the correct regression value.

(b) The special case that $\varepsilon = 0$ results in a Huber smoothed absolute loss.

Fig. 1. Huber smoothed ε-insensitive loss. On the horizontal axis is $y_i - \langle \beta, x_i \rangle$ while the vertical axis plots f_ε in blue, and $\frac{\partial f_\varepsilon}{\partial \beta}$ in red. In both plots $h = \frac{1}{2}$.

Table 1. Accuracies for each method and regularizer. See text for the experimental setting. The k-support norm achieved higher acuracies on average for all loss functions.

	f_2	f_{2-}	f_h	$f_{\log}$	$f_{\exp}$	f_{abs}	f_ε
$\|\beta\|_k^{sp}$	0.883 ± 0.058	0.883 ± 0.058	0.890 ± 0.057	0.889 ± 0.056	0.888 ± 0.060	0.889 ± 0.065	0.886 ± 0.062
$\|\beta\|_1$	0.870 ± 0.062	0.870 ± 0.062	0.868 ± 0.069	0.872 ± 0.063	0.876 ± 0.065	0.870 ± 0.077	0.879 ± 0.059
$\|\beta\|_2$	0.871 ± 0.071	0.871 ± 0.071	0.872 ± 0.065	0.872 ± 0.066	0.870 ± 0.067	0.867 ± 0.071	0.872 ± 0.063

differentiability.

$$f_\varepsilon(\beta, X, y) = \sum_{i=1}^n \begin{cases} 0 & \text{if } y_i - \langle \beta, x_i \rangle > -\varepsilon + h \\ \frac{(y_i - \langle \beta, x_i \rangle + \varepsilon - h)^2}{4h} & \text{if } |y_i - \langle \beta, x_i \rangle + \varepsilon| \leq h \\ -y_i + \langle \beta, x_i \rangle - \varepsilon & \text{if } y_i - \langle \beta, x_i \rangle < -\varepsilon - h \end{cases} \tag{23}$$

$$+ \sum_{i=1}^n \begin{cases} 0 & \text{if } y_i - \langle \beta, x_i \rangle < \varepsilon - h \\ \frac{(y_i - \langle \beta, x_i \rangle - \varepsilon + h)^2}{4h} & \text{if } |y_i - \langle \beta, x_i \rangle - \varepsilon| \leq h \\ y_i - \langle \beta, x_i \rangle - \varepsilon & \text{if } y_i - \langle \beta, x_i \rangle > \varepsilon + h \end{cases}$$

$$\frac{\partial f_\varepsilon}{\partial \beta} = \sum_{i=1}^n \begin{cases} 0 & \text{if } y_i - \langle \beta, x_i \rangle > -\varepsilon + h \\ \frac{\langle \beta, x_i \rangle x_i + (-\varepsilon + h - y_i) x_i}{2h} & \text{if } |y_i - \langle \beta, x_i \rangle + \varepsilon| \leq h \\ x_i & \text{if } y_i - \langle \beta, x_i \rangle < -\varepsilon - h \end{cases} \tag{24}$$

$$+ \sum_{i=1}^n \begin{cases} 0 & \text{if } y_i - \langle \beta, x_i \rangle < \varepsilon - h \\ \frac{\langle \beta, x_i \rangle x_i + (\varepsilon - h - y_i) x_i}{2h} & \text{if } |y_i - \langle \beta, x_i \rangle - \varepsilon| \leq h \\ -x_i & \text{if } y_i - \langle \beta, x_i \rangle > \varepsilon + h \end{cases}$$

$$L_\varepsilon = \frac{\gamma}{h} \tag{25}$$

Here we have decomposed the ε insensitive loss into two hinge components to emphasize the relationship to Huber smoothed hinge loss (cf. Section 4). A plot of the loss and its gradient is shown in Figure 1. In the case that $\varepsilon = 0$ we get a Huber smoothed absolute loss function as a special case (denoted f_{abs} in the sequel), and the curvature of the loss function at $y_i - \langle \beta, x_i \rangle = 0$ is doubled, therefore the Lipschitz constant is double that of the one sided hinge loss.

In the case that $k = d$, we recover the special case of ε-support vector regression (ε-SVR) [10]. If we set $k = 1$ we get an ℓ_1 regularized variant of ε-SVR. In the case that $\varepsilon = 0$ this ℓ_1 regularized variant is equivalent to regularized least absolute deviations regression [15]. In Equation (23), $\varepsilon < 0$ is equivalent to $\varepsilon > 0$ but with a constant value added to the loss everywhere, i.e. the minimizer is the same.

8 Experiments

We have applied each of the algorithms above to a toy classification problem conceptually similar to that reported in [1]. In all cases, we perform model selection for $k \in \{1, \dots, d\}$ and $\lambda = 10^i$, $i \in \{-15, \dots, 5\}$. We compare additionally to

Table 2. Mean squared errors (MSE) for each method and regularizer. See text for the experimental setting. f_2, f_{2-}, f_{abs}, and f_ε achieved the lowest MSEs with the k-support norm regularizer giving best results on average.

	f_2	f_{2-}	f_h	$f_{\log}$	$f_{\exp}$	f_{abs}	f_ε
$\|\beta\|_k^{sp}$	$1.21e2 \pm 4.89e1$	$1.21e2 \pm 4.89e1$	$1.78e2 \pm 1.00e2$	$3.33e3 \pm 5.39e3$	$1.59e3 \pm 2.89e3$	$1.25e2 \pm 5.41e1$	$2.21e2 \pm 1.51e1$
$\|\beta\|_1$	$1.25e2 \pm 4.81e1$	$1.25e2 \pm 4.81e1$	$2.21e2 \pm 9.63e1$	$1.13e4 \pm 9.89e3$	$6.16e3 \pm 4.82e3$	$1.48e2 \pm 1.76e2$	$2.16e2 \pm 1.66e1$
$\|\beta\|_2$	$1.49e2 \pm 4.75e1$	$1.49e2 \pm 4.74e1$	$1.81e2 \pm 7.66e1$	$4.18e3 \pm 8.00e3$	$3.08e3 \pm 4.88e3$	$1.50e2 \pm 5.34e1$	$2.25e2 \pm 1.56e1$

the special fixed cases $k = 1$ and $k = d$ corresponding to ℓ_1 and ℓ_2 regularization, respectively.

Output labels were generated randomly with equal probability. The first 15 dimensions were set by multiplying the label by a fixed vector of 15 samples from a zero mean Gaussian and adding Gaussian noise (i.e. a noisy signal is contained in the first 15 dimensions). The subsequent 50 dimensions were set to zero mean Gaussian noise (i.e. the subsequent dimensions contain no signal and should be ignored). 50 samples were used for training, 50 for validation, and 250 for testing. Table 1 gives the mean accuracies for each method across 20 random problem instances, while Table 2 gives the mean squared error (MSE). For ε-insensitive loss we arbitrarily set $\varepsilon = 1$. For all methods with a Huber smothing parameter, we set $h = \frac{1}{10}$.

It should be noted that several of the methods employed here for classification were developed for regression (squared loss, absolute loss, and ε-insensitive loss). The experiments performed here were done primarily to validate their correct implementation.

9 Conclusions

We have described and implemented a large number of loss functions for non-differentiably regularized risk optimization with proximal splitting methods. These loss functions in combination with the k-support norm yield a large number of learning algorithms proposed in the literature as special cases. Assuming zero model error, each of these loss functions is sufficient to yield a statistically consistent algorithm[2] (provided regularization goes to zero at a sufficient rate as the number of samples goes to infinity) [3, Theorem 4]. However, their finite sample behavior varies substantially. We hope that their implementation and description in a common framework will facilitate their analysis and employment in machine learning studies and applications.

References

1. Argyriou, A., Foygel, R., Srebro, N.: Sparse prediction with the k-support norm. In: NIPS. pp. 1466–1474 (2012)

[2] Huber smoothed ε-insensitive loss requires that $\varepsilon - h < 1$ for consistency in the binary classification setting, $y_i \in \{-1, +1\}$.

2. Argyriou, A., Micchelli, C.A., Pontil, M.: When is there a representer theorem? vector versus matrix regularizers. Journal of Machine Learning Research 10, 2507–2529 (2009)
3. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. Journal of the American Statistical Association 101(473), 138–156 (2006)
4. Chapelle, O.: Training a support vector machine in the primal. Neural Computation 19(5), 1155–1178 (May 2007)
5. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. 20(3), 273–297 (Sep 1995)
6. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: ICML. pp. 148–156 (1996)
7. Gkirtzou, K., Honorio, J., Samaras, D., Goldstein, R., Blaschko, M.B.: fMRI analysis of cocaine addiction using k-support sparsity. In: ISBI (2013)
8. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction. Springer Series in Statistics, Springer (2009)
9. Ng, A.Y.: Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In: Proceedings of the twenty-first international conference on Machine learning (2004)
10. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA (2001)
11. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B) 58, 267–288 (1996)
12. Tikhonov, A.: Solution of incorrectly formulated problems and the regularization method. In: Soviet Math. Doklady. vol. 4, pp. 1035–1038 (1963)
13. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)
14. Wang, L., Zhu, J., Zou, H.: The doubly regularized support vector machine. Statistica Sinica 16(2), 589–616 (2006)
15. Wang, L., Gordon, M.D., Zhu, J.: Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In: Proceedings of the Sixth International Conference on Data Mining. pp. 690–700 (2006)
16. Zhu, J., Rosset, S., Hastie, T., Tibshirani, R.: 1-norm support vector machines. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA (2004)
17. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67, 301–320 (2005)