Convex envelopes of complexity controlling penalties: the case against premature envelopment

Vladimir Jojic
Stanford University

Suchi Saria
Stanford University

Daphne Koller
Stanford University

Abstract

Convex envelopes of the cardinality and rank function, l_1 and nuclear norm, have gained immense popularity due to their sparsity inducing properties. This has given rise to a natural approach to building objectives with sparse optima whereby such convex penalties are added to another objective. Such a heuristic approach to objective building does not always work. For example, addition of an L_1 penalty to the KL-divergence fails to induce any sparsity, as the L_1 norm of any vector in a simplex is a constant. However, a convex envelope of KL and a cardinality penalty can be obtained that indeed trades off sparsity and KL-divergence.

We consider the cases of two composite penalties, elastic net and fused lasso, which combine multiple desiderata. In both of these cases, we show that a hard objective relaxed to obtain penalties can be more tightly approximated. Further, by construction, it is impossible to get a better convex approximation than the ones we derive. Thus, constructing a joint envelope across different parts of the objective provides a means to trade off tightness and computational cost.

1 Introduction

Compact summarization of data succinctly describes many tasks shared across areas such as statistics, machine learning, information theory, and computational biology. The quality of a model is measured as a trade-off between reconstruction error and the complexity of the model's parametrization. A number of costs exist that capture this trade-off: MDL (Barron et al., 1998), BIC (Schwarz, 1978), AIC (Akaike, 1973), Bayes factor (Kass and Raftery, 1995) and so on. These costs are discontinuous and in some cases even their evaluation (e.g., Bayes factors) can be challenging.

Recent work on discovering compact representations of data has focused on combinations of convex losses, mostly stemming from generalized linear models, and convex sparsity promoting penalties. The uniqueness of optima in these costs combined with parameter sparsity makes them quite desirable. The joint simplicity of generalized linear losses and sparsity of parameters makes the models easily interpretable, a critical feature when the results of such fits need to be communicated across fields.

The sparsity inducing penalties have found applications across a variety of disciplines: ℓ_1, most commonly under the guise of lasso(Tibshirani, 1996) and elastic net (Zou and Hastie, 2005); nuclear norm within a variety of applications such as compressed sensing (Candes et al., 2006); and recommender matrix reconstruction (Candes and Recht, 2009).

However, the crucial observation that ℓ_1 is the tightest convex relaxation of cardinality and nuclear norm for rank (Fazel, 2002) has not been leveraged. At the same time, the family of lasso penalties continues to grow via juxtaposition. The constituent weights of these fragmented penalties are adjusted via cross validation posing serious computational problems, and at the same time eroding interpretability. The goal of this paper is to illustrate how the juxtaposition of convex penalties corresponds to piecemeal relaxation of a hard penalty and to show that the provably tightest convex approximation can be constructed for hard penalties.

We also show that in some fairly common cases, elastic net and fused lasso, tighter convex relaxations exist. In other cases, we show a dramatic failure of the juxtaposition approach in inducing any change to the objective: the sparsifying convex penalty fails to sparsify.

We show how an envelope can be computed and

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

envelope-penalized losses can be optimized. We illustrate how these tasks can be accomplished both when an envelope is available in closed form and in cases where a closed form is not available.

In contrast to recent work on relationships between submodular set functions, convex envelopes of these functions, and sparsity inducing norms (Bach, 2010), we focus on envelopes of composite penalties involving cardinality. The key insight is that a joint envelope over multiple hard penalties is tighter than envelopment of each penalty in isolation.

The structure of the paper is as follows: first we state results on Fenchel duality and envelopes and show the main steps in deriving a convex envelope of cardinality. We then give two simple optimization algorithms for cases where the closed-form of the envelope is not available. The first algorithm numerically evaluates an envelope. The second algorithm provide a general blueprint for optimizing envelope penalized losses. We proceed to illustrate how simple juxtaposition of penalties, KL-divergence and ℓ_1 in this case, can fail to produce any effect, but also show the existence of sparsified KL-divergence. We then focus on the elastic net penalty, composed of ℓ_2 and ℓ_1. For this penalty, a relaxation of cardinality and ℓ_2, we show the tightest convex relaxation and compare the performance of these two relaxations in the tasks of support recovery. Given a closed-form solution of the envelope, we show a simple coordinate descent method for a loss penalized by this envelope. Finally, we look at another example of a composite penalty, fused lasso, and show that it also has a corresponding hard penalty, related to its degrees of freedom, and use its envelope to illustrate its benefits in the context of a biological application of copy number variation analysis.

2 Convex envelope

Following Hiriart-Urruty and Lemarchal (1993) we will assume that the functions that we aim to envelope are not identically $+\infty$ and can be minorized by an affine function on a set of interest, for example box $[-1,1]^n$. Here we restate the definition of a conjugate and the theorem about the tightness of the envelope.

Definition The conjugate of a function $f(x \in \mathcal{X})$ is the function

$$f^*(y) = \sup_{x \in \mathcal{X}} \langle y, x \rangle - f(x). \tag{1}$$

Theorem 2.1 *For a function $f(x \in \mathcal{X})$ the envelope (biconjugate)*

$$f^{**}(z) = (f^*)^*(z) = \sup_y \langle y, z \rangle - f^*(y) \tag{2}$$

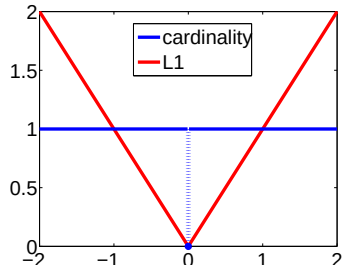

Figure 1: ℓ_1 is an envelope of $\mathbf{card}(\cdot)$ only in the $[-1,1]$ interval

is the pointwise supremum of all the affine functions on $\mathbf{R}^n$ majorized by f.

Hence the envelope of f is the tightest convex underapproximate of the function f.

2.1 An example of convex envelope derivation: Cardinality to ℓ_1

We show the computation of the convex envelope for $\mathbf{card}(\cdot)$ as a prototype for such derivations (Boyd and Vandenberghe, 2004). The penalty is given as

$$f_\lambda(x \in [-1,1]^n) = \lambda \mathbf{card}(x)$$

and its conjugate is by definition

$$f^*(y \in \mathbf{R}^n) = \sup_{x \in [-1,1]^n} \langle x, y \rangle - \lambda \mathbf{card}(x).$$

We will assume an ordering σ of coordinates y such that $y^2_{\sigma(i)} \geq y^2_{\sigma(i-1)}$ and introduce an auxiliary variable r

$$f^*(y) = \max_r \sup_{x \in [-1,1]^r} \left\langle x, y_{\sigma(1:r)} \right\rangle - \lambda r$$

Once r is fixed, the supremum is achieved when the inner product is maximized and given a constraint that $x \in [-1,1]^n$, that is $x_i = \operatorname{sgn}(y_i)$, hence

$$f^*(y) = \max_r \sum_{i=1}^{r} \left(\left| y_{\sigma(i)} \right| - \lambda \right) = \sum_{i=1}^{n} \left(|y_i| - \lambda \right)_+ .$$

The biconjugate is

$$f^{**}(z \in [-1,1]^n) = \sup_{y \in \mathbf{R}^n} \langle z, y \rangle - \sum_{i=1}^{n} \left(|y_i| - \lambda \right)_+$$

Since the above supremum is separable in coordinates, we can reason about each coordinate in isolation and, for each, consider cases when the positive part is zero and non-negative. Elementary reasoning produced the closed form solution of the supremum

$$\sum_{i=1}^{n} \lambda \left| z_i \right|$$

Hence the biconjugate of $\lambda\mathbf{card}(\cdot)$ on $[-1,1]^n$ is $\lambda\|\cdot\|_1$. Figure 1 shows both the f and its envelope f^{**}. Note that the envelope is only valid on the box $[-C,C]^n$. WLOG we will assume that $C=1$ in the rest of the paper. As the box boundaries grow, the envelope f^{**} can be shown to tend to a constant. However, a change in the box boundaries corresponds to scaling of the penalty parameter λ.

3 Evaluating the envelope and optimizing envelope-penalized losses

In this section we show numerical methods for two common tasks: pointwise evaluation of the envelope and optimization of the envelope-penalized losses. Pointwise evaluation of the envelope is useful for exploring the behavior of the envelope, such as the level sets' shape. However, for the second task of minimizing penalized losses, explicit pointwise evaluation of the envelope can be bypassed in the interest of efficient optimization.

3.1 Pointwise evaluation of the envelope

We focus on envelopes of functions that do not have a closed form solution but turn out to be numerically tractable. The requirement in this scheme is that pointwise computation of the conjugate $f^*(y) = \sup_x \langle y,x\rangle - f(x)$ is tractable. Given this assumption, we can find an x_y^* that achieves the supremum, where the subscript is used to indicate dependence on y. The envelope is given as

$$f^{**}(z) = \sup_y h(y,z), \quad h(y,z) = \langle y,z\rangle - f^*(y). \tag{3}$$

If the conjugate has a closed form then $\nabla_y h(y,z)$ can be computed symbolically. Alternatively, we can apply Danskin's theorem (Bertsekas, 1999) to obtain $\nabla_y h(y,z)$. For this, we need to define the set X_y that contains all x such that they achieve the supremum of $\langle y,x\rangle - f(x)$. Thus,

$$\nabla_y h(y,z) = \{z - x^* : x^* \in X_y\}. \tag{4}$$

In either case, we can evaluate the envelope using the standard subgradient method, for example Algorithm 1. We note that the form of the envelope problem makes it amenable to the smoothing methods of (Nesterov, 2005).

However, the conjugate is not always tractable. An example of such a function is

$$f(\beta) = \|D - M\beta\|_2^2 + \lambda\mathbf{card}(\beta)$$

a linear regression problem with target variable $D \in \mathbf{R}^m$ and predictor matrix $M \in \mathbf{R}^{m\times n}$. Evaluating $f^*(y) = \sup_\beta \langle \beta, y\rangle - \lambda f(\beta)$ in this case is just as hard as the original problem, subset selection. This is not surprising as the tightness of the convex envelope implies that it touches the global minimum of the function. Hence, envelopment does not erase the hardness of the original problem.

Algorithm 1 A subgradient method for computing envelope $f^{**}(z)$

Input: z
Output: $f^{**}(z)$
  for $k=1$ to MAXITER **do**
    $y^{k+1} = y^k + (1/k)\nabla h(y,z)$
  end for
  return $\langle y^{\text{MAXITER}+1}, z\rangle - f^*(y^{\text{MAXITER}+1})$

3.2 The variational inequality approach to minimizing envelope-penalized losses

Minimization of an envelope-penalized loss

$$\text{PLoss}(z) = \text{Loss}(z) + f^{**}(z) \tag{5}$$

seems to require pointwise-evaluation of the envelope. When a closed-form envelope is available, this does not pose a problem. In the previous section, we have shown a method for evaluating a convex envelope numerically. It would seem that in order to minimize Eq.5 we would need to wrap an outer loop optimizing z around an inner loop computing the envelope. To further complicate matters, the approximate nature of the inner loop would then necessitate an appeal to approximate subgradient schemes such as (Solodov and Svaiter, 2000). We bypass these considerations by recasting the optimization problem as a variational inequality problem.

The problem of minimizing the cost in Eq. 5 can be written as a saddle problem

$$\inf_z \sup_y \text{Loss}(z) + h(y,z) \tag{6}$$

which is equivalent to finding $V = \begin{bmatrix} y \\ z \end{bmatrix}$ such that

$$\langle F(V), W - V\rangle \geq 0, \ \forall W \in \mathcal{V}$$

where

$$F\left(\begin{bmatrix} y \\ z \end{bmatrix}\right) = \begin{bmatrix} -\nabla_y h(y,z) \\ \nabla_z \text{Loss}(z) + y \end{bmatrix} \tag{7}$$

and $\mathcal{V} = \mathbf{R}^n \times \mathbf{B}$, with $\mathbf{B}$ denoting the set on which f^{**} is defined, for example box $[-1,1]^n$. Algorithm 2 solves the variational inequality problem (Nemirovski, 2005; Tseng, 2008). One choice for $D(V,W)$

is $(1/2)\|V-W\|_2^2$, but other distance terms can be used, especially on coordinates that correspond to set **B**. Finally, L denotes the Lipschitz constant of F and is influenced primarily by the Loss function.

Algorithm 2 Proximal point algorithm for a VIP

for $k=1$ to MAXITER **do**
  $W^k = \text{argmin}_{V\in\mathcal{V}} \left\{\left\langle V, F(V^k)\right\rangle + L\cdot D(V,V^k)\right\}$
  $V^{k+1} = \text{argmin}_{V\in\mathcal{V}} \left\{\left\langle V, F(W^k)\right\rangle + L\cdot D(V,V^k)\right\}$
end for
return $V_z^{\text{MAXITER}+1}$

This numerical algorithm requires $O(L/\epsilon)$ iterations to achieve an ϵ-approximate solution (Nemirovski, 2005; Tseng, 2008). This complexity still compares favorably to projected subgradients optimizing a closed form cost, which in general require $O(1/\epsilon^2)$ iterations for a solution of the same quality.

4 Sparsified KL-divergence

KL-divergence's use is ubiquitous in machine learning, and more recent work has investigated incorporating penalties and constraints with the divergence (Graça et al., 2007). Obtaining sparse multinomials has been attempted by using a negative-weight Dirichlet prior (Bicego et al., 2007), as well as by inducing sparsity in sufficient statistics for mixture proportions via l_∞ norm (Graca et al., 2009). The approach of using L_1 norm to induce sparsity has served well in sparsifying objectives. However, in the case of KL this augmentation fails immediately and obviously as soon as the optimization problem is written

$$\begin{array}{ll}\text{minimize} & \left\langle q, \log\frac{q}{p}\right\rangle + \lambda\|q\|_1 \\ \text{subject to} & q_i \geq 0, \forall_i \\ & \sum_i q_i = 1,\end{array}$$

since the q is constrained to be in a simplex, its L_1 norm is always 1, thus the penalty is a constant. Of course, this does not mean that sparsification of KL divergence is impossible.

Recalling that the L_1 norm is the tightest convex relaxation (envelope) of **card**$(\cdot)$, the de-facto sparsifying penalty, we can pose a different problem

$$\begin{array}{ll}\text{minimize} & f(q) = \left\langle q, \log\frac{q}{p}\right\rangle + \lambda\mathbf{card}(q) \\ \text{subject to} & q_i \geq 0, \forall_i \\ & \sum_i q_i = 1,\end{array}$$

and seek the convex envelope of f on the n-dimensional simplex Δ_n. We obtain the conjugate

$$f^*(y) = \max_{r\in\{1,..,n\}} \left\{\log\left\{\sum_{i=1}^{r}\exp\left\{y_{\sigma(i)}\right\}p_{\sigma(i)}\right\} - \lambda r\right\} \quad (8)$$

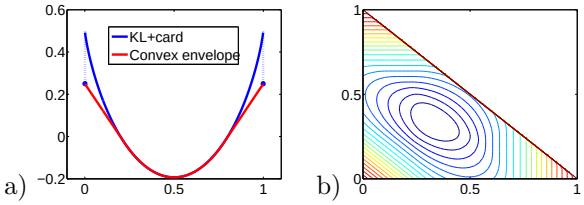

Figure 2: a) KL$(p||\text{Unif}) + 0.33\mathbf{card}(p)$ and its convex envelope in 2d simplex. b) contours of the convex envelope of KL$(p||\text{Unif}) + 0.33\mathbf{card}(p)$ in 3d simplex. The contours become closer to straight lines closer to low complexity solutions, the corners and sides of the simplex. This shape of contours is reminiscent of ℓ_1's contours. The switch to entropy-like curved contours occurs towards the middle of the simplex.

where σ is an order such that

$$\exp\left\{y_{\sigma(i)}\right\}p_{\sigma(i)} \geq \exp\left\{y_{\sigma(i+1)}\right\}p_{\sigma(i+1)}, \forall i.$$

The derivation of the envelope can proceed from this point by conjugating $f^*(y)$. Here we opt to use the scheme proposed in Sec. 3 to numerically evaluate the envelope. The requirements for application of that scheme were that we can compute the conjugate and its gradient efficiently. Exact computation of $f^*(y)$ requires sorting of a vector of length n to obtain the order σ, and a pass through the logarithm of cumulative sums to obtain an r^* that maximizes Eq. 8. We can plug in f^* into the definition of $h(y,z)$ in Eq. 3 and obtain a subgradient $\nabla_y h(y,z)$

$$\frac{\partial h(y,z)}{\partial y_{\sigma(i)}} = \begin{cases} z_{\sigma(i)}, & \sigma(i) > r^* \\ z_{\sigma(i)} - \frac{\exp\left\{y_{\sigma(i)}p_{\sigma(i)}\right\}}{\sum_{j=1}^{r^*}\exp\left\{y_{\sigma(j)}p_{\sigma(j)}\right\}}, & \sigma(i) \leq r^* \end{cases}$$

Figure 2 shows a convex envelope of KL$(p||\text{Unif}) + 0.33\mathbf{card}(p)$ in 2d simplex as well as its contour plots in 3d simplex. Here we can conclude that a joint envelope of the KL-divergence and cardinality yields a sparsifying objective, whereas a separate envelope of the objective parts yields only KL-divergence. Not only does the simple addition of penalties yield a loose convex approximation of the desired objective, but in some cases it fails completely to produce any effect. Thus, a joint envelope of all parts of the objective is preferable when computationally feasible.

5 Envelope of cardinality and ℓ_2

Elastic net penalty (Zou and Hastie, 2005) consists of a sum of ℓ_2 and ℓ_1. These two penalties induce a grouping effect, i.e. joint selection of correlated predictors, and sparsity in regression weights. Considering again

ℓ_1 as a convex relaxation of cardinality, we can ask if jointly enveloping a penalty that combines ℓ_2 and cardinality would yield a different convex penalty. Indeed, it does, and the resulting penalty is the "Berhu" penalty (Owen, 2006). The intuition behind the introduction of this penalty was the desire to produce a robust ridge regression estimator, and the form of the penalty was arrived at by rearranging parts of the Huber penalty. Our construction shows that this penalty is the tightest convex relaxation of cardinality and ℓ_2.

Envelope of $\ell_2 + \mathbf{card}(\cdot)$ The penalty is

$$f(x) = \frac{\kappa}{2}\|x\|_2^2 + \lambda \mathbf{card}(x) \tag{9}$$

where $x \in B_n = [-1,1]^n$ and

$$\mathbf{card}(x) = \sum_{i=1}^{n} [x_i \neq 0]. \tag{10}$$

We can extend f on the whole $\mathbf{R}^n$ so that the conjugate is well defined

$$f_\lambda(x) = \begin{cases} \infty, & x \notin B_n \\ \frac{\kappa}{2}\|x\|_2^2 + \lambda \mathbf{card}(x), & x \in B_n \end{cases} \tag{11}$$

However, this does not affect computation of the conjugate, so we can focus on the $x \in B_n$.

Conjugate The conjugate of f is given by

$$f^*(y) = \sup_{x \in B_n} \langle x, y\rangle - \frac{\kappa}{2}\|x\|_2^2 - \lambda \mathbf{card}(x) \tag{12}$$

and rewriting

$$f^*(y) = \frac{1}{2\kappa}\|y\|_2^2 + \sup_{x \in B_n} -\frac{\kappa}{2}\left\|x - \frac{1}{\kappa}y\right\|_2^2 - \lambda \mathbf{card}(x) \tag{13}$$

The conjugate simplifies into a coordinate-wise representation:

$$f_i^*(y) = \begin{cases} 0 & y_i^2 \leq 2\lambda\kappa \\ \frac{1}{2\kappa}y_i^2 - \lambda, & 2\lambda\kappa \leq y_i^2 \leq \kappa^2 \\ |y_i| - \lambda - \frac{\kappa}{2}, & \kappa^2 \leq y_i^2 \end{cases} \tag{14}$$

Biconjugate We form the biconjugate which also remains in coordinate-wise form:

$$f^{**}(z \in B_n) = \sup_{y \in \mathbf{R}^n} \sum_i y_i z_i - f_i^*(y) \tag{15}$$

Since the supremum is separable, we can express it in a pointwise manner

$$f_i^{**}(z_i) = \sup_{y_i} \begin{cases} z_i y_i, & y_i^2 \leq 2\lambda\kappa \\ z_i y_i - \frac{1}{2\kappa}y_i^2 + \lambda, & 2\lambda\kappa \leq y_i^2 \leq \kappa^2 \\ z_i y_i + \frac{\kappa}{2} - |y_i| + \lambda, & \kappa^2 \leq y_i^2 \end{cases} \tag{16}$$

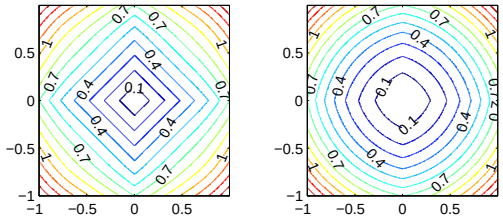

Figure 3: Two relaxations of $0.2\mathbf{card}(x) + 0.5\ell_2(x)$ on the $[-1,1]^2$ box. The convex envelope (left) and elastic net penalty $0.2\ell_1(x) + 0.5\ell_2(x)$ (right). The level sets of the convex envelope for the same value of penalty are smaller as a consequence of the tightness of the envelope compared to elastic net. Note that the contours of the convex envelope close to low complexity solutions are composed of straight lines, with an ℓ_1-like diamond for small coordinates. The regime switch to the curved, ℓ_2-like curves occurs with larger coordinates.

We note that the function under the supremum is continuous at the boundaries of the conditions above $y_i^2 = 2\lambda\kappa$ and $y_i^2 = \kappa^2$. Computing optima for each of the three conditions, we obtain the envelope below. Note that the third condition in Eq. 16 is never active.

$$f_i^{**}(z_i) = \begin{cases} |z_i|\sqrt{2\lambda\kappa}, & |z_i| \leq \sqrt{\frac{2\lambda}{\kappa}} \\ \frac{\kappa}{2}z_i^2 + \lambda, & |z_i| \geq \sqrt{\frac{2\lambda}{\kappa}} \end{cases} \tag{17}$$

and

$$f^{**}(z) = \sum_i f_i^{**}(z_i)$$

Optimizing penalty and loss jointly We can now take the envelope f^{**} of the target function f and add it to the loss to obtain a penalized loss $\mathrm{PL}(z)$. WLOG, we assume that all predictors are standardized, i.e. $\sum_j X_{i,j} = 0$ and $X_i^T X_i = 1$.

$$\mathrm{PL}(z) = (1/2)\|Y - Xz\|_2^2 + \sum_i \begin{cases} |z_i|\sqrt{2\lambda\kappa}, & |z_i| \leq \sqrt{\frac{2\lambda}{\kappa}} \\ \frac{\kappa}{2}z_i^2 + \lambda, & |z_i| \geq \sqrt{\frac{2\lambda}{\kappa}} \end{cases}$$

This problem can be optimized by coordinate descent. To derive the update equations, we will use the following notation: $Y_{-i} = Y - \sum_{j \neq i} X_j z_j$.

Hence coordinate descent iterates the following update, with $Z_{-i} = Y_{-i}^T X_i$

$$z_i = \begin{cases} 0, & |Z_{-i}| \leq \sqrt{2\lambda\kappa} \\ Z_{-i} - \mathrm{sgn}(Z_{-i})\sqrt{2\lambda\kappa}, & \frac{|Z_{-i}|}{1+\kappa} \leq \sqrt{\frac{2\lambda}{\kappa}} \\ \frac{1}{1+\kappa}Z_{-i}, & \frac{|Z_{-i}|}{1+\kappa} \geq \sqrt{\frac{2\lambda}{\kappa}} \end{cases}$$

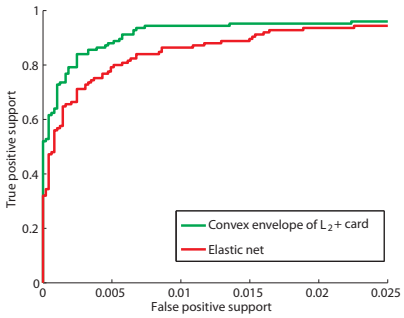

Figure 4: Comparison of elastic net and the convex envelope of ℓ_2+**card**$(\cdot)$ on a synthetic support recovery task. The data was generated from a linear model with 5000 predictors with true support consisting of 125 predictors. The support consisted of 25 groups of 5 correlated predictors, with correlation bounded from below by 0.9. For both methods, λ_1 corresponded to a penalty associated with a sparsity component and λ_2 corresponded to squared ℓ_2 components. We scanned a grid of values for both λ_1 and λ_2 ranging from -15 to 0 in log domain, in increments of 1. The optimal penalties are chosen via 5-fold cross-validation. We show an ROC curve on the task of support recovery for the best performing choice of λs for each penalty. The false positive range is truncated at 2.5%, reflecting the imbalance between the sizes of the true positive and the true negative sets in this task.

As a consequence of Theorem 2.1 the derived penalty is the tightest convex relaxation of **card**$(\cdot) + \ell_2$. Figure 3 illustrates the differences between the elastic net penalty and the envelope of **card**$(\cdot) + \ell_2$. The envelope has smaller level sets for the same penalty levels, which is a direct consequence of its tightness compared to elastic net. Further, the envelope smoothly transitions from the ℓ_1-like diamond contours for small coordinates to ℓ_2-like round curves for larger coordinates.

In Figure 4 we are showing a comparison of perfomance of the two penalties in the task of support recovery. The penalty constants λ_1 and λ_2 corresponding to the sparsity and ℓ_2 portions were sought on the grid of values, independently for the two methods. The ROC curves illustrate the best recovery rates achieved using the two penalties with their respective parameters set by 5-fold cross-validation. The convex envelope penalty recovers higher a proportion of the support compared to the elastic net penalty.

6 Envelope of non-zero-block penalty

The fused lasso penalty proposed in (Tibshirani et al., 2005) is meant to combine sparsity and smoothness. Sparsity in a vector of values is induced by ℓ_1 on the coordinates. Smoothness is induced by ℓ_1 on the difference between subsequent coordinate values. This penalty, in the case of a neighborhood relationship induced by a chain, is

$$\sum_i \lambda_1 |x_i| + \sum_i \lambda_2 |x_i - x_{i-1}|$$

combined with a squared error loss

$$\sum_i \|y - x\|_2^2 \tag{18}$$

where y (the data vector) and x are both column vectors. The combination of this squared loss and fused lasso penalty yields the fused lasso signal approximator.

One characterization of the complexity of the fused lasso fit is given by the number of distinct non zero blocks (Eq. 13 (Tibshirani et al., 2005)). While optima of the fused lasso cost exhibit a "blocky" structure, the penalty can deviate significantly from simply counting the number of non-zero blocks. Here we show how the convex envelope of the non-zero-block count function can be optimized, even if the closed form is not available.

6.1 Conjugate of degrees of freedom

The complexity measure of a fused lasso fit, degrees-of-freedom, stated in Equation 13 (Tibshirani et al., 2005), is the cardinality of the complement of the set $\{i : x_i \neq 0\} \cup \{i : x_i - x_{i-1} = 0, x_j, x_{j-1} \neq 0\}$ which is exactly $\{i : x_i \neq 0, x_i = x_{i-1}\}$. Hence we can formulate the hard penalty relaxed by the fused lasso as

$$\textbf{card}(\langle x \neq 0, x \neq x_\pi \rangle) = \lambda \sum_i (x_i \neq 0, x_i \neq x_{i-1})$$

where x_π denotes neighbors of x. We define the nonzero-block penalty as

$$f_{nzb}(x) = \lambda \langle x \neq 0, x \neq x_\pi \rangle .$$

The conjugate is then given by

$$f^*_{nzb}(y) = \sup_{x \in [-1,1]^N} \langle x, y \rangle - \lambda \langle x \neq 0, x \neq x_\pi \rangle . \tag{19}$$

Next, we show that the optimization of the supremum can be achieved by a Viterbi-like algorithm over a small number of hidden states, by characterizing the optimal assignments to x.

Proposition 6.1 *The supremum of* $h(x) = \langle x, y \rangle - \lambda \langle x \neq 0, x \neq x_\pi \rangle$ *where* $x \in B_n$ *can be achieved by* $x \in S_n = \{-1, 0, 1\}^n$.

Proof Assume that there exists a maximizing assignment for $x^* \in [-1,1]^N$. The assignment x^* can be broken down into blocks of consecutive indices $I_1, \dots I_m$ such that $x^*_{I_j} = a_j$, a slight abuse of notation, and $a_{j-1} \neq a_j$ and at least one $a_j \notin \{-1,0,1\}$. We can now write the function $h(x^*) = \sum_{j=1}^m h_j(x_{I_j})$ where $h_j(x_{I_j}) = \langle x_{I_j}, y_{I_j} \rangle - \lambda$. We now proceed to construct x' such that $h(x') \geq h(x^*)$. For each block I_j, such that $a_j \in \{-1,0,1\}$, we keep the same values $x'_{I_j} = a_j$, hence $h_j(x'_{I_j}) \geq h_j(x^*_{I_j})$. For blocks such that $a_j \notin \{-1,0,1\}$ and $y_{I_j} = 0$ setting $x'_{I_j} = \text{sgn}\left(\sum_{l \in I_j} y_l\right)$ achieves $h_j(x'_{I_j}) \geq h_j(x^*)$. Hence, we can obtain an x' from x^* such that $h(x') \geq h(x^*)$ where $x' \in \{-1,0,1\}^N$ since x^* achieves a supremum so does x'.

Hence, we can reformulate computation of the conjugate as optimization in this smaller discrete set:

$$f^*_{nzb}(y) = \max_{x \in S_n} \langle x, y \rangle - \lambda \langle x \neq 0, x \neq x_\pi \rangle \qquad (20)$$

and highlight the applicability of a Viterbi-like algorithm in computing the conjugate

$$\begin{aligned} f^*_{nzb}(y) = \quad & \max_{x_1 \in S_1} x_1 y_1 - \lambda(x_1 \neq 0) + \\ & \max_{x_2 \in S_1} x_2 y_2 - \lambda(x_2 \neq 0, x_2 \neq x_1) + \dots \\ & \max_{x_n \in S_1} x_n y_n - \lambda(x_n \neq 0, x_n \neq x_{n-1}). \end{aligned} \qquad (21)$$

Given a maximizing assignment x^* and using Eq. 4, we can compute a subgradient of $\nabla_y h(y,z) = z - x^*$ and use it in Algorithm 1, specialized to the problem of the envelope of the non-zero-block penalty. The complexity of the Viterbi pass is linear in the length of the data vector and hence computation of the subgradient is linear.

6.2 Optimization of a squared loss penalized by the envelope of non-zero-block count

We now specialize the two updates in Algorithm 2 to the problem of optimizing the loss from Eq. 18 penalized by the envelope of non-zero-block count

$$(1/2)\left\|y - z\right\|_2^2 + \lambda f^{**}_{nzb}(z). \qquad (22)$$

Adopting notation Viterbi$(\cdot,\cdot)$ to indicate Viterbi decoding for the cost given in Eq. 21 we can specify the updates inside the loop of Algorithm 2

$$\begin{aligned} W^k_y &= V^k_y - \frac{1}{L}\left(-V^k(z) + \text{Viterbi}(V^k_y, \lambda)\right) \\ W^k_z &= \Pi_{B_n}\left[V^k_z - \frac{1}{L}\left(V^k(z) - d\right) + V^k_y\right] \\ V^{k+1}_y &= V^k_y - \frac{1}{L}\left(-W^k(z) + \text{Viterbi}(W^k_y, \lambda)\right) \\ V^{k+1}_z &= \Pi_{B_n}\left[W^k_z - \frac{1}{L}\left(W^k(z) - d\right) + V^k_y\right]. \end{aligned}$$

The complexity of an iteration of this algorithm is linear in the size of the data sample.

6.3 Comparison to fused lasso

For purposes of comparison, we use the loss from Eq.18,

$$L(z) = (1/2)\left\|y - z\right\|_2^2$$

and compare the performance of the two penalties, fused lasso and the envelope of non-zero-block counter f^{**}_{nzb}.

From the HapMap CNV Project we obtained copy number estimates for chromosomes 12,13,15 and 18 of individual NA10855. For both methods we perform pathwise fit, controlling for the number of degrees-of-freedom and plotting the best reconstruction error for a model of that complexity, see Figure 5 . This regime corresponds to a purely unsupervised task of compression. In these examples, envelope penalized squared-error loss can yield models that capture more of the data variance with a smaller number of degrees-of-freedom. Notably, these examples have significant amounts of structure. In cases of simpler signals, near constant or with a small number of switching points, the two penalties do not recover noticeably different models.

7 Conclusion

Convex envelopes of cardinality offer a direct way to capture parameter complexity and have yielded two very useful penalties, ℓ_1 and nuclear norm. In spite of the introduction of novel lasso style penalties, the principled approach to combining these penalties has not been put forward. We illustrated the hazards of simple composition of penalties on the example of KL-divergence and ℓ_1. In that case, we show that introduction of the sparsifying penalty fails to yield any sparsity. However, the joint convex envelope of KL-divergence and cardinality yields the desired sparsified objective. Further, we show that two well known penalties, elastic net and fused lasso, can be viewed as relaxations of hard penalties and that those penalties can in turn be enveloped to yield even tighter convex relaxations. We also show two simple algorithms for envelope evaluation and optimization of envelope penalized losses. We suggest that convex envelopes with guarantees on tightness provide a powerful approach to combining penalties, explicitly trading off computational costs, and investigating novel hybrid penalties.

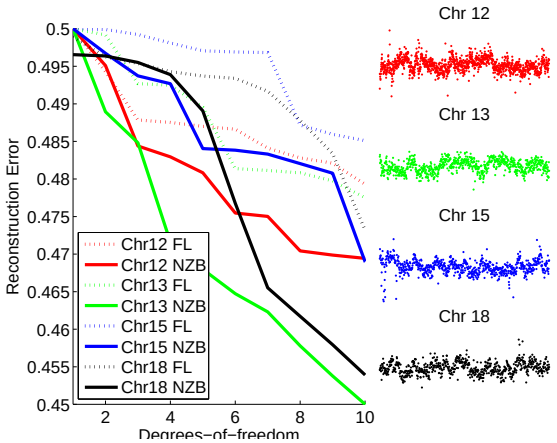

Figure 5: Model fits to copy number variation data from the HapMap project. The data is shown on the right. The measurements span four chromosomes from individual NA10855. The plot shows the reconstruction error for varying complexity of models. In the plot, FL denotes fused lasso penalized loss and NZB denotes loss penalized by the envelope of non-zero-block counts.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. 2nd internat. Sympos. Inform. Theory, Tsahkadsor 1971, 267-281 (1973)., 1973.

F. Bach. Structured sparsity-inducing norms through submodular functions. In *NIPS*, 2010.

A. Barron, J. Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *Information Theory, IEEE Transactions on*, 44(6): 2743 –2760, October 1998.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, September 1999. ISBN 1886529000.

M. Bicego, M. Cristani, and V. Murino. Sparseness achievement in hidden markov models. In *ICIAP '07: Proceedings of the 14th International Conference on Image Analysis and Processing*, pages 67–72, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2877-5.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787.

E J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489 – 509, Feb. 2006.

M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Dept. of Elec. Eng., Stanford University, 2002.

J. Graca, K. Ganchev, B. Taskar, and F. Pereira. Posterior vs parameter sparsity in latent variable models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 664–672, 2009.

J. Graça, K. Ganchev, and B. Taskar. Expectation maximization and posterior constraints. In *NIPS*, 2007.

J.-B. Hiriart-Urruty and C. Lemarchal. *Convex analysis and minimization algorithms. Part 1: Fundamentals*. Springer-Verlag, 1993.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):pp. 773–795, 1995.

A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. on Optimization*, 15(1):229–251, 2005. ISSN 1052-6234.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005. ISSN 0025-5610.

A. B. Owen. A robust hybrid of lasso and ridge regression. Technical report, Stanford University, 2006.

G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, 1978.

M. V. Solodov and B. F. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of bregman functions. *Mathematics of Operations Research*, 25:214–230, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108, 2005.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization Optim.*, 2008.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.