Screening Data Points in Empirical Risk Minimization with Ellipsoidal Regions and Safe Loss Functions

Grégoire Mialon, Alexandre d'Aspremont and Julien Mairal

Context

- To screen := running a "simple" test to discard useless variables in a data set before running an optimization algorithm.
- Seminal work by [El Ghaoui et al., 2010] for the Lasso. From KKT conditions and geometry of the Lasso, design a screening test consisting in checking an inequality on the dual variable for a set containing the optimal dual variable.
- Applications: memory gains; dynamic rules [Fercoq et al., 2015] (screening performed as the optimization algorithm proceeds) speeding up convergence.
- The test can be problem-specific or generic.
- Scarce litterature for sample screening.

We propose a new, generic way to design tests for sample screening.

Context

In supervised learning, the goal is to learn a prediction function h given labeled training data $(a_i, b_i)_{i=1,\dots,n}$ with $a_i \in \mathbb{R}^p$, and $b_i \in \mathbb{R}$:

$$\min_{h \in \mathcal{H}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} f(h(a_i), b_i)}_{\text{Empirical risk, data fit}} + \underbrace{\lambda R(h)}_{\text{Regularization}} .$$

In many applications, f is convex and h is linear, i.e. $h(a_i) = x^\top a_i$ (in what follows, we do not use an intercept without loss of generality).

Context

By introducing the **margin** t defined as $t_i = x^\top a_i - b_i$ (regression) or $t = b_i x^\top a_i$ (classification), the problem becomes

$$\min_{x \in \mathbb{R}^p, t \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} f(t_i) + \lambda R(x) \qquad (1)$$
$$\text{s.t } t = \mathbf{diag}(b) A x,$$

with

$$f(t) = \begin{cases} \max(1-t, 0) & \text{(SVMs)} \\ \log(\exp^{-t} + 1) & \text{(Logistic Regression)}, \end{cases} \quad R(x) = \begin{cases} \frac{1}{2}\|x\|_2^2 & \text{in general,} \\ \|x\|_1 & \text{for inducing sparsity,} \end{cases}$$

and many others...

Margins and Safe loss functions

Definition (Safe loss function)

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a continuous convex loss function such that $\inf_{t \in \mathbb{R}} \varphi(t) = 0$. We say that φ is a safe loss if there exists a non-singleton and non-empty interval $\mathcal{I} \subset \mathbb{R}$ such that

$$t \in \mathcal{I} \implies \varphi(t) = 0.$$

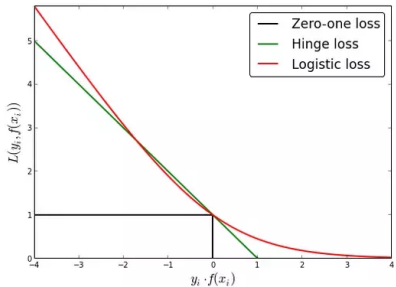


Figure 1: Example. The Hinge loss admits a flat area while the Logistic loss does not.

Losses with a flat area and dual sparsity

A dual problem (obtained from Lagrange duality) to the ERM (1) above is

$$\max_{\nu \in \mathbb{R}^n} D(\nu) = \frac{1}{n} \sum_{i=1}^{n} -f_i^*(\nu_i) - \lambda R^* \left(-\frac{A^T \nu}{\lambda n} \right).$$

At the optimum, $x^\star = -\frac{A^\top \nu^\star}{\lambda n}$.

Lemma (Safe loss and dual sparsity)

Consider the primal dual problems above. Denoting by $x^\star$ and $\nu^\star$ the optimal primal and dual variables respectively, we have for all $i = 1, \ldots, n$,

$$\nu_i^\star \in \partial f_i(a_i^\top x^\star).$$

Consequence: For both classification and regression, the sparsity of the dual solution is related to loss functions that have flat regions.

Safe screening rule for data points

Theorem (Safe rules for data points)

For a loss having a flat region $\mathcal{I}$, consider a subset $\mathcal{X}$ containing the optimal solution $x^\star$. If, for a given data point (a_i, b_i), the margin $t \in \mathring{\mathcal{I}}$ for all x in $\mathcal{X}$, where $\mathring{\mathcal{I}}$ is the interior of $\mathcal{I}$, then this data point can be discarded from the dataset.

We assume that there exists $\mu > 0$ such that $\mathcal{I} = [-\mu, \mu]$ for safe regression losses and $\mathcal{I} = [\mu, +\infty)$ for classification.

Consequence: If $\max_{x \in \mathcal{X}} |a_i^\top x - b_i| < \mu$ (regression) or $\min_{x \in \mathcal{X}} b_i a_i^\top x > \mu$ (classification), with $\mathcal{X}$ a set which is known to contain $x^\star$, then a_i can be discarded from the data set A (or "screened").

Safe screening rule

Question: How to find a good set $\mathcal{X}$?

Safe screening rule

Question: How to find a good set $\mathcal{X}$?

- It has to be small.

Safe screening rule

Question: How to find a good set $\mathcal{X}$?

- It has to be small.
- It has to be tractable.

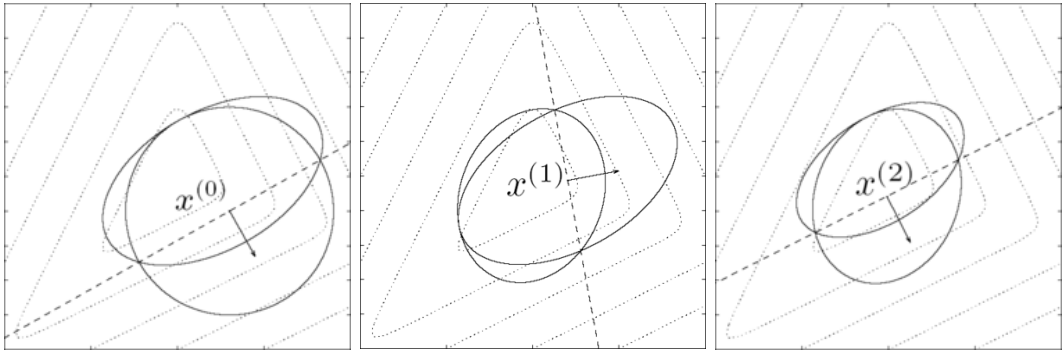
Safe screening rule

Question: How to find a good set $\mathcal{X}$?

- It has to be small.
- It has to be tractable.

$\min_{x \in \mathcal{X}} b_i a_i^\top x$ and $\max_{x \in \mathcal{X}} |a_i^\top x - b_i|$ are closed form when $\mathcal{X}$ is an ellipsoid!

Finding $\mathcal{X}$: Ellipsoid Method (Nemirovski and Yudin, 1976)



Step 0. Step 1. Step 2.

Wrapping up

Algorithm 1 Building ellipsoidal test regions

initialization: Given $\mathcal{E}^0(x_0, E_0)$ containing $x^\star$;
while $k < nb_{\text{steps}}$ **do**
    • Compute a gradient g of the objective in x_k;
    • $\tilde{g} \leftarrow g / \sqrt{g^T E_k g}$;
    • $x_{k+1} \leftarrow x_k - \frac{1}{p+1} E_k \tilde{g}$;
    • $E_{k+1} \leftarrow \frac{p^2}{p^2-1}(E_k - \frac{2}{p+1} E_k \tilde{g} \tilde{g}^T E_k)$;
For classification problems:
for each sample a_i in A **do**
    if $\min b_i x^\top a_i \geq \mu$ for $x \in \mathcal{E}^{nb_{\text{steps}}}$ **then**
        Discard a_i from A.

Comparison to other safe regions

- [Ogawa et al., 2013] : pathwise computation properties of SVM.
- [Shibagaki et al., 2016] : when the objective is strongly convex, $x^\star \in \mathcal{B}(x, \frac{2\Delta(x)}{\lambda})$ with x a current iterate and $\Delta(x)$ a duality gap of the problem.

Table 1: State-of-the-art comparison for sample screening

Method	Strongly convex	Non strongly convex	Generic
Pathwise SVM [Ogawa et al., 2013]	✓	✗	✗
Duality Gap [Shibagaki et al., 2016]	✓	✗	✓
Ellipsoid (Ours)	✓	✓	✓

Building safe losses

Logistic loss $f(t) = \log(1 + e^{-t})$ and $\Omega(x) = -x \log(-x) + \mu|x|$ for $x \in [-1, 0]$. We have $\Omega^*(y) = -e^{y+\mu-1}$. Convolving Ω^* with f yields

$$f_\mu(x) = \begin{cases} e^{x+\mu-1} - (x+\mu) & \text{if } x + \mu - 1 \leq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Smooth and asymptotically robust. The entropic part of Ω makes this penalty strongly convex hence f_μ is smooth [Nesterov, 2005]. Finally, the ℓ_1 penalty ensures that the dual is sparse thus making the screening usable.

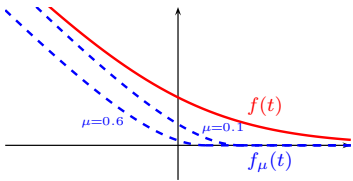


Figure 2: Classification loss.

Experiments: many data sets contain useless samples

- In many datasets, there are a lot of samples to screen.
- *MNIST* ($n = 60{,}000$) and *SVHN* ($n = 604{,}388$) both represent digits, encoded by using the output of a two-layer convolutional kernel network [Mairal, 2016] leading to feature dimension $p = 2304$. *RCV-1* ($n = 781{,}265$) represents sparse TF-IDF vectors of categorized newswire stories ($p = 47{,}236$).

Table 2: Percentage of samples that can be discarded for problems trained with an ℓ_1-Safe Logistic loss.

Dataset	MNIST	SVHN	RCV-1
$\lambda = 10^{-3}$	0 %	2 %	12 %
$\lambda = 10^{-4}$	27 %	17 %	42 %
$\lambda = 10^{-5}$	65 %	54 %	75 %

Experiments: a trade-off between screening and optimizing

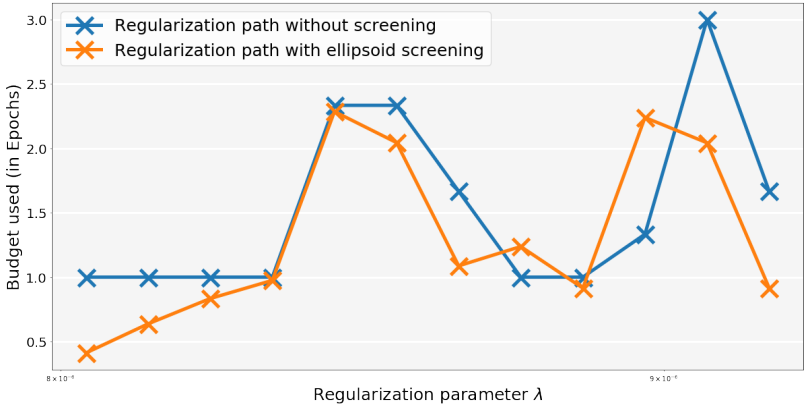


Figure 3: Regularization path of a Squared Hinge SVM trained on MNIST. Computational budget used vs. Regularization parameter in the path. Screening enables computational gains.

References I

El Ghaoui, L., Viallon, V., and Rabbani, T. (2010). Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems. *arXiv e-prints*, page arXiv:1009.4219.

Fercoq, O., Gramfort, A., and Salmon, J. (2015). Mind the duality gap: safer rules for the Lasso. In *International Conference on Machine Learning (ICML)*.

Mairal, J. (2016). End-to-end kernel learning with supervised convolutional kernel networks. In *Advance in Neural Information Processing Systems (NIPS)*.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152.

Ogawa, K., Suzuki, Y., and Takeuchi, I. (2013). Safe screening of non-support vectors in pathwise svm computation. In *International Conference on Machine Learning (ICML)*.

References II

Shibagaki, A., Karasuyama, M., Hatano, K., and Takeuchi, I. (2016). Simultaneous Safe Screening of Features and Samples in Doubly Sparse Modeling. In *International Conference on Machine Learning (ICML)*.