Screening Data Points in Empirical Risk Minimization

Grégoire Mialon

Inria Paris

Gatsby Unit, December 16, 2019

Publication

Julien Mairal

Alexandre d'Aspremont

- G. Mialon, A. d'Aspremont, J. Mairal : Screening Data Points in Empirical Risk Minimization via Ellipsoidal Regions and Safe Loss Functions preprint arXiv:1912.02566. 2019.

Context

- To screen := To automatically discard useless variables before running an optimization algorithm.
- Seminal work by [El Ghaoui et al., 2010] for the Lasso. From KKT conditions, if a dual optimal variable satisfies a given inequality constraint, the corresponding primal optimal variable must be zero. Check this on a set which contains the optimal dual variable.
- Applications : memory gains ; dynamic rules [Fercoq et al., 2015] (screening performed as the optimization algorithm proceeds) speeding up convergence.
- Scarce litterature for sample screening.

Context

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

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

In most applications, 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 by $t = x^\top a_i - b_i$ (regression) or $t = b_i x^\top a_i$ (classification), the problem becomes

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

with

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

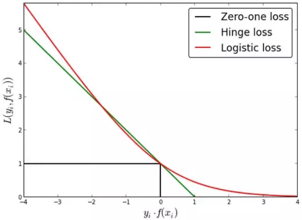
and many others...

Margins

Definition (Safe loss function)

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a continuous convex loss function such that $\inf_{t \in \mathbb{R}} \phi(t) = 0$. 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

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



The Hinge loss admits a flat area while the Logistic loss does not.

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").

Losses with a flat area and dual sparsity

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

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

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.

Proof

We consider the dual problem (obtained from Lagrange duality)

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

We always have $P(x) \geq D(\nu)$. Since there exists a pair (x, t) such that $Ax = t$ (Slater's conditions), we have $P(x^\star) = D(\nu^\star)$ and $x^\star = -\frac{A^\top \nu^\star}{\lambda n}$ at the optimum.

From the definition of safe loss functions and assuming that $b_i a_i^\top x \in \mathring{\mathcal{I}}$, f_i is differentiable at $a_i^\top x^\star$ with $\nu_i^\star = f_i'(a_i^\top x^\star) = 0$. ∎

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.

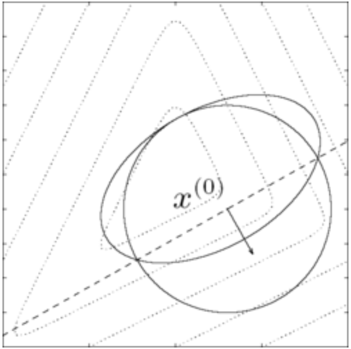
Safe screening rule

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

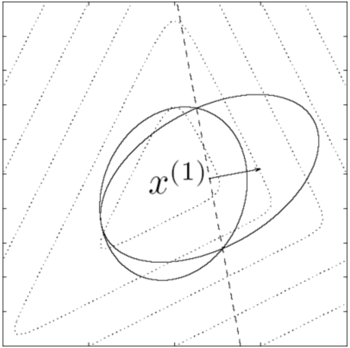
- It has to be small.
- It has to be tractable.

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

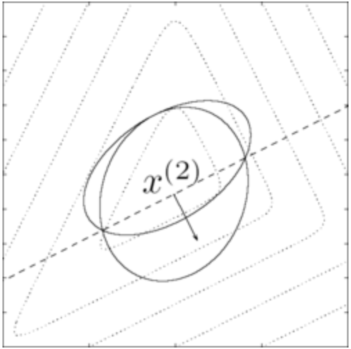
Finding $\mathcal{X}$: Ellipsoid Method (Nemirovski and Yudin, 1976)



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 obj 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.

	Strongly convex	Non strongly convex	Generic
Pathwise SVM	✓	✗	✗
Duality Gap	✓	✗	✓
Ellipsoid	✓	✓	✓

State of the art for sample screening

Building safe losses

When the ERM problem does not admit a sparse dual solution, safe screening is not possible.

Definition (Infimum convolution)

Let $f : \mathbb{R}^p \to \mathbb{R} \cup \{-\infty, +\infty\}$ be an extended real-valued function and Ω a convex term. Let f_μ be defined as

$$f_\mu = \min_{z \in \mathbb{R}^p} f(z) + \mu \Omega^* \left(\frac{t - z}{\mu} \right). \tag{1}$$

f_μ is called the infimum convolution of f and Ω^*, which may be written as $f \,\square\, \Omega^*$.

Note that f_μ is convex as the minimum of a convex function in (t, z). We recover the Moreau-Yosida smoothing [Moreau, 1962, Yosida, 1980] and its generalization when Ω is respectively a quadratic term or a strongly-convex term [Nesterov, 2005].

Building safe losses

Lemma (Regularized dual for classification)

Consider the modified classification problem

$$\min_{x \in \mathbb{R}^p, t \in \mathbb{R}^n} f_\mu(t) + \lambda R(x) \quad \text{s.t.} \quad t = \mathbf{diag}(b)Ax. \tag{$\mathcal{P}_2'$}$$

The dual of $\mathcal{P}_2'$ is

$$\max_{\nu \in \mathbb{R}^n} -f^*(-\nu) - \lambda R^* \left(\frac{A^T \mathbf{diag}(b)\nu}{\lambda} \right) - \mu\Omega(-\nu). \tag{2}$$

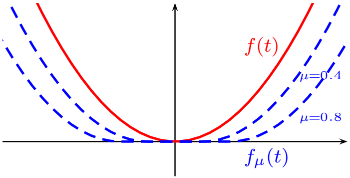
It will be possible, in many cases, to induce sparsity in the dual if Ω is the ℓ_1-norm, or another sparsity-inducing penalty.

Building safe losses

Quadratic loss $f: t \mapsto \|t\|^2/2$ and $\Omega(x) = \|x\|_1$. Then $\Omega^*(y) = \mathbf{1}_{\|y\|_\infty \leq 1}$ (see *e.g.* [Bach et al., 2012]), and

$$f_\mu(t) = \sum_{i=1}^{n} \frac{1}{2}[|t_i| - \mu]_+^2. \tag{3}$$

The parameter μ encourages the loss to be flat (it is exactly 0 when $\|t\|_\infty \leq \mu$).



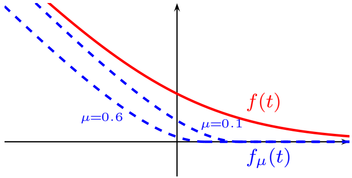
Regression loss.

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

$$f_\mu(x) = \begin{cases} e^{x+\mu-1} - (x+\mu) & \text{if } x + \mu - 1 \leq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Smooth and asymptotically robust.



Classification loss.

Experiments

- 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 dimensions $p = 2304$. *RCV-1* ($n = 781{,}265$) represents sparse TF-IDF vectors of categorized newswire stories ($p = 47{,}236$).

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

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

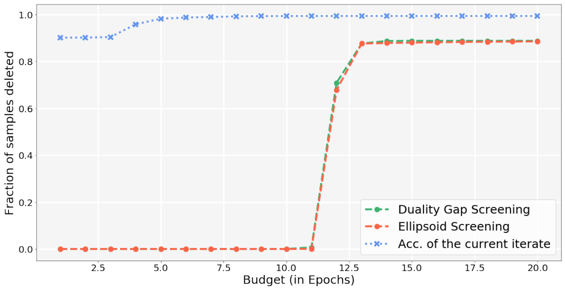
Experiments

Dataset	MNIST	SVHN
$\lambda = 1.0$	89 / 89	87 / 87
$\lambda = 10^{-1}$	95 / 95	91 / 91
$\lambda = 10^{-2}$	98 / 98	90 / 92
$\lambda = 10^{-3}$	34 / 50	0 / 0

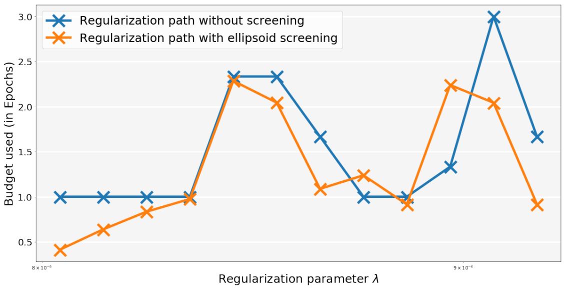
Dataset	RCV-1
$\lambda = 1$	85 / 85
$\lambda = 10$	80 / 80
$\lambda = 100$	68 / 68

Percentage of samples screened in an ℓ_2 penalized SVM with Squared Hinge loss (Ellipsoid (ours) / Duality Gap) given the epochs made at initialization.

Experiments

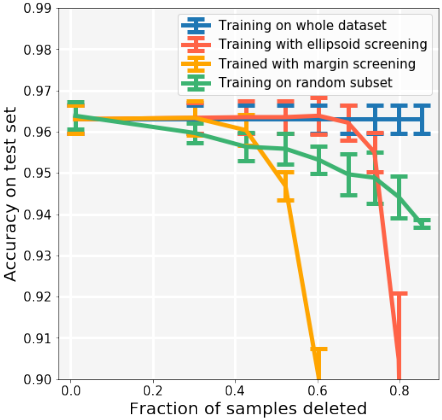


Fraction of samples discarded vs Epochs done for two screening strategies along with test accuracy of the current iterate (ℓ_2-Squared Hinge, MNIST).

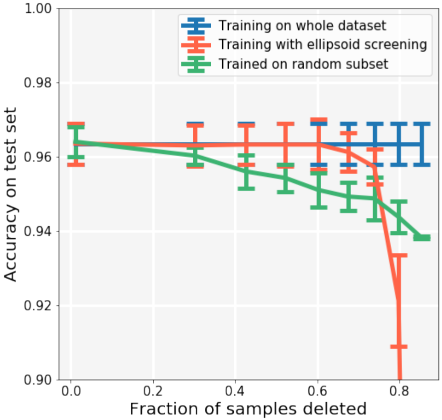


Regularization path of a Squared Hinge SVM trained on MNIST. Screening enables computational gains.

Experiments (Proof of concept)



(a) SVHN and ℓ_2 Squared Hinge



(b) SVHN and ℓ_1 Safe Logistic

Dataset compression in classification.

References I

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106.

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)*.

Moreau, J.-J. (1962). Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A MAth*.

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

References II

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)*.

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)*.

Yosida, K. (1980). Functional analysis. *Berlin-Heidelberg*.