A Proofs.

A.1 Proof of Lemma 2.4

Proof. At the optimum,

$$P(x^*) - D(\nu^*) = \frac{1}{n}\sum_{i=1}^{n} f_i(a_i^\top x) + f_i^*(\nu_i) + \lambda R(x) + \lambda R^*\left(-\frac{A^T\nu}{\lambda n}\right) = 0.$$

Adding the null term $\langle x, -\frac{A^\top \nu}{n}\rangle - \langle x, -\frac{A^\top \nu}{n}\rangle$ gives

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{f_i(a_i^\top x) + f_i^*(\nu_i) - a_i^\top x \nu_i}_{\geq 0} + \underbrace{\lambda\left(R(x) + R^*\left(-\frac{A^\top\nu}{\lambda n}\right) - \left\langle x, -\frac{A^\top\nu}{\lambda n}\right\rangle\right)}_{\geq 0} = 0,$$

since Fenchel-Young's inequality states that each term is greater or equal to zero. We have a null sum of non-negative terms; hence, each one of them is equal to zero. We therefore have for each $i = 1 \dots n$:

$$f(a_i^\top x) + f^*(\nu_i) = a_i^\top x \nu_i,$$

which corresponds to the equality case in Fenchel-Young's relation, which is equivalent to $\nu_i^* \in \partial f_i(a_i^\top x^*)$. ■

A.2 Proof of Lemma 3.3

Proof. The Lagrangian of the problem writes:

$$L(x, \nu, \gamma) = a_i^\top x - b_i + \nu\left(1 - (x-z)^T E^{-1}(x-z)\right) - \gamma g^T(x-z),$$

with $\nu, \gamma \geq 0$. When maximizing in x, we get:

$$\frac{\partial L}{\partial x} = a_i + 2\nu(E^{-1}z - E^{-1}x) - \gamma = 0.$$

We have $\nu > 0$ since the opposite leads to a contradiction. This yields $x = z + \frac{1}{2\nu}(Ea_i - \gamma Eg)$ and $(x-z)^T E^{-1}(x-z) = 1$ at the optimum which gives $\nu = \frac{1}{2}\sqrt{(a_i-\gamma)^T E(a_i-\gamma)}$.

Now, we have to minimize

$$g(\nu, \gamma) = a_i\left(z + \frac{1}{2\nu}(Ea_i - \gamma Eg)\right) - \gamma^\top\left(\frac{1}{2\nu}(Ea_i - \gamma Eg)\right).$$

To do that, we consider the optimality condition

$$\frac{\partial g}{\partial \gamma} = -\frac{1}{2\nu}a_i Eg - \frac{1}{2\nu}g^T E a_i + \frac{\gamma}{\nu}g^T E g = 0,$$

which yields $\gamma = \frac{g^T E a_i}{g^T E g}$. If $g^T E a_i < 0$ then $\gamma = 0$ in order to avoid a contradiction.

In summary, either $g^T E a_i \leq 0$ hence the maximum is attained in $x = z + \frac{1}{2\nu}Ea_i$ and is equal to $a_i z + \sqrt{a_i^T E a_i} - y_i$, or $g^T E a_i > 0$ and the maximum is attained in $x = z + \frac{1}{2\nu}E(a_i - \gamma Eg)$ and is equal to $a_i\left(z + \frac{1}{2\nu}E(a_i - \gamma g)\right) - b_i$ with $\nu = \frac{1}{2}\sqrt{(a_i-\gamma)^T E(a_i-\gamma)}$ and $\gamma = \frac{g^T E a_i}{g^T E g}$. ■

A.3 Proof of Lemma 4.1

Proof. We can write $\mathcal{P}_1'$ as

$$\begin{array}{ll} \text{minimize} & \tilde{f}(\tilde{x}) + \lambda\tilde{R}(\tilde{x}) \\ \text{subject to} & \tilde{A}\tilde{x} = -b \end{array} \tag{10}$$

in the variable $\tilde{x} = (t, x) \in \mathbb{R}^{n+p}$ with $\tilde{f}\colon \tilde{x} \mapsto f_\mu(t)$ and $\tilde{R}\colon \tilde{x} \mapsto R(x)$ and $\tilde{A} \in \mathbb{R}^{n\times(n+p)} = (\mathrm{Id}, -A)$. Since the constraints are linear, we can directly express the dual of this problem in terms of the Fenchel conjugate of the objective (see *e.g.* Boyd and Vandenberghe (2004), 5.1.6). Let us note $f_0 = \tilde{f} + \lambda\tilde{R}$. For all $y \in \mathbb{R}^{n+p}$, we have

$$\begin{aligned} f_0^*(y) &= \sup_{x\in\mathbb{R}^{n+p}} \langle x, y\rangle - \tilde{f}(x) - \lambda\tilde{R}(x) \\ &= \sup_{x_1\in\mathbb{R}^n, x_2\in\mathbb{R}^p} \langle x_1, y_1\rangle + \langle x_2, y_2\rangle - f(x_1) - \lambda R(x_2) \\ &= f_\mu^*(y_1) + \lambda R^*\left(\frac{y_2}{\lambda}\right). \end{aligned}$$

It is known from Beck and Teboulle (2012) that $f_\mu = f \,\square\, \Omega_\mu^* = (f^* + \Omega_\mu^{**})^*$ with $\Omega_\mu^* = \mu\Omega^*(\frac{\cdot}{\mu})$. Clearly, $\Omega_\mu^{**} = \mu\Omega$. If Ω is proper, convex and lower semicontinuous, then $\Omega = \Omega^{**}$. As a consequence, $f_\mu^* = (f^* + \mu\Omega)^{**}$. If $f^* + \mu\Omega$ is proper, convex and lower semicontinuous, then $f_\mu^* = f^* + \mu\Omega$, hence

$$f_0^*(y) = f^*(y_1) + \lambda R^*\left(\frac{y_2}{\lambda}\right) + \mu\Omega(y_1).$$

Now we can form the dual of $\mathcal{P}_1'$ by writing

$$\text{maximize} \quad -\langle -b, \nu\rangle - f_0^*(-\tilde{A}^T\nu) \tag{11}$$

in the variable $\nu \in \mathbb{R}^n$. Since $-\tilde{A}^T\nu = (-\nu, A^T\nu)$ with $\nu \in \mathbb{R}^n$ the dual variable associated to the equality constraints,

$$f_0^*(-\tilde{A}^T\nu) = f^*(-\nu) + \lambda R^*\left(\frac{A^T\nu}{\lambda}\right) + \mu\Omega(-\nu).$$

Injecting f_0^* in the problem and setting ν instead of $-\nu$ (we optimize in $\mathbb{R}$) concludes the proof. ■

A.4 Lemma A.1

Lemma A.1 (Bounding f_μ)**.** *If $\mu \geq 0$ and Ω is a norm then*

$$f(t) - \delta(t) \leq f_\mu(t) \leq f(t), \quad \textit{for all } t \in \mathrm{dom} f$$

with $\delta(t) = \max_{\|\frac{u}{\mu}\|^* \leq 1} g^T u$ *and* $g \in \partial f(t)$.

Proof. If Ω is a norm, then $\Omega(0) = 0$ and Ω^* is the indicator function of the dual norm of Ω hence non-negative. Moreover, if $\mu > 0$ then, $\forall z \in \text{dom} f$ and $\forall t \in \mathbb{R}^n$,

$$f_\mu(t) \leq f(z) + \mu \Omega^* \left(\frac{t-z}{\mu} \right).$$

In particular, we can take $t = z$ hence the right-hand inequality. On the other hand,

$$\begin{aligned} f_\mu(t) - f(t) &= \min_z f(z) + \mu I_{\|\frac{z-t}{\mu}\|^* \leq 1} - f(t) \\ &= \min_{\|\frac{u}{\mu}\|^* \leq 1} f(t+u) - f(t). \end{aligned}$$

Since f is convex,

$$f(t+u) - f(t) \geq g^T u \text{ with } g \in \partial f(t).$$

As a consequence,

$$f_\mu(t) - f(t) \geq \min_{\|\frac{u}{\mu}\|^* \leq 1} g^T u.$$

■

A.5 Proof of Lemma 4.4

Proof. The proof is trivial given the inequalities in Lemma A.1. ■

A.6 Proof of Screening-friendly regression

Proof. The Fenchel conjugate of a norm is the indicator function of the unit ball of its dual norm, the ℓ_∞ ball here. Hence the infimum convolution to solve

$$f_\mu(x) = \min_{z \in \mathbb{R}^n} \{ f(z) + \mathbf{1}_{\|x-z\|_\infty \leq \mu} \} \tag{12}$$

Since $f(x) = \frac{1}{2n}\|x\|_2^2$,

$$f_\mu(x) = \min_{z \in \mathbb{R}^n} \frac{1}{2n} z^T z + \mathbf{1}_{\|x-z\|_\infty \leq \mu}.$$

If we consider the change of variable $t = x - z$, we get:

$$f_\mu(x) = \min_{t \in \mathbb{R}^n} \frac{1}{2n} \|x-t\|_2^2 + \mathbf{1}_{\|t\|_\infty \leq \mu}.$$

The solution t^* to this problem is exactly the proximal operator for the indicator function of the infinity ball applied to x. It has a closed form

$$\begin{aligned} t^* &= \text{prox}_{\mathbf{1}_{\|\cdot\|_\infty \leq \mu}}(x) \\ &= x - \text{prox}_{\left(\mathbf{1}_{\|\cdot\|_\infty \leq \mu}\right)^*}(x), \end{aligned}$$

using Moreau decomposition. We therefore have

$$t^* = x - \text{prox}_{\mu\|\cdot\|_1}(x).$$

Hence,

$$f_\mu(x) = \frac{1}{2n}\|x - t^*\|_2^2 = \frac{1}{2n}\|\text{prox}_{\mu\|\cdot\|_1}(x)\|_2^2.$$

But, $\text{prox}_{\mu\|\cdot\|_1}(t) = \text{sgn}(t) \times [|t| - \mu]_+$ for $t \in \mathbb{R}$, where $[x]_+ = \max(x, 0)$. ■

B Additional examples.

Squared hinge loss. Let us consider a problem with a quadratic loss $f \colon t \mapsto \|1 - t\|_2^2/2$ designed for a classification problem, and consider $\Omega(x) = \|x\|_1 + \mathbf{1}_{x \succeq 0}$. We have $\Omega^*(y) = \mathbf{1}_{y \succeq -1}$, and

$$f_\mu(t) = \sum_{i=1}^n [1 - t_i - \mu, 0]_+^2,$$

which is a squared Hinge Loss with a threshold parameter μ and $[.]_+ = \max(0, .)$.

C Additional experimental results.

Reproducibility. The data sets did not require any pre-processing except *MNIST* and *SVHN* on which exhaustive details can be found in Mairal (2016). For both regression and classification, the examples were allocated to train and test sets using scikit-learn's *train-test-split* (80% of the data allocated to the train set). The experiments were run three to ten times (depending on the cost of the computations) and our error bars reflect the standard deviation. For each fraction of points deleted, we fit three to five estimators on the screened dataset and the random subset before averaging the corresponding scores. The optimal parameters for the linear models were found using a simple grid-search.

Accuracy of our safe logistic loss. The accuracies of the Safe Logistic loss we build is similar to the accuracies obtained with the Squared Hinge and the Logistic losses on the datasets we use in this paper thus making it a realistic loss function.

RCV-1. Table 4 shows additional screening results on RCV-1 with a ℓ_2 penalized Squared Hinge loss SVM.

Epochs	10	20
$\lambda = 1$	7 / 84	85 / 85
$\lambda = 10$	80 / 80	80 / 80
$\lambda = 100$	68 / 68	68 / 68

Table 4: RCV-1 : Percentage of samples screened in an ℓ_2 penalized SVM with Squared Hinge loss (Ellipsoid (ours) / Duality Gap) given the epochs made at initialization.

Dataset	MNIST	SVHN	RCV-1
Logistic + ℓ_1	0.997 (0.01)	0.99 (0.0003)	0.975 (1.0)
Logistic + ℓ_2	0.997 (0.001)	0.99 (0.0003)	0.975 (1.0)
Safelog + ℓ_1	0.996 (0.0)	0.989 (0.0)	0.974 (1e-05)
Safelog + ℓ_2	0.996 (0.0)	0.989 (0.0)	0.975 (1e-05)
Squared Hinge + ℓ_1	0.997 (0.03)	0.99 (0.03)	0.975 (1.0)
Squared Hinge + ℓ_2	0.997 (0.003)	0.99 (0.003)	0.974 (1.0)

Table 3: Averaged best accuracies on test set (best λ in a Logarithmic grid from $\lambda = 0.00001$ to 1.0).

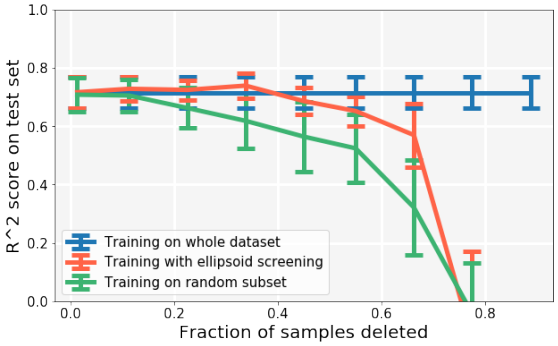
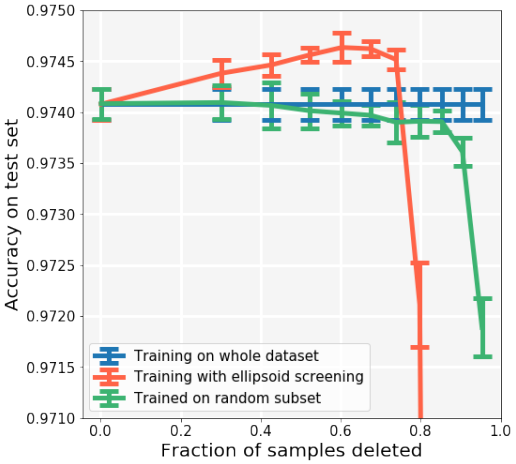
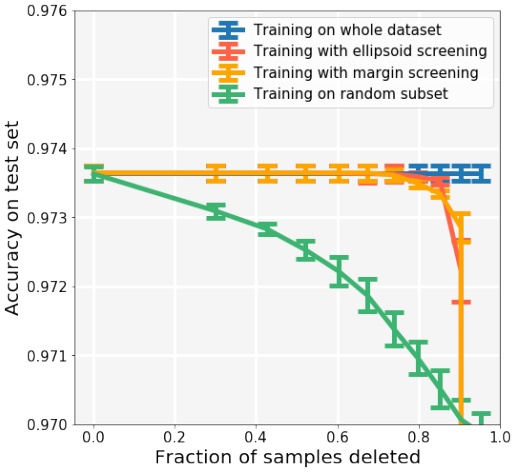


Figure 7: Dataset compression for the Lasso trained on a synthetic dataset. The scores given by the screening yield a ranking that is better than random subsampling.

Lasso regression. The Lasso objective combines an ℓ_2 loss with an ℓ_1 penalty. Since its dual is not sparse, we will instead apply the safe rules offered by the screening-friendly regression loss (7) derived in Section 4.3 and illustrated in Figure 2, combined with an ℓ_1 penalty. We can draw an interesting parallel with the SVM, which is naturally sparse in data points. At the optimum, the solution of the SVM can be expressed in terms of data points (the so-called support vectors) that are close to the classification boundary, that is the points that are *the most difficult* to classify. Our screening rule yields the analog for regression: the points that are easy to predict, i.e. that are close to the regression curve, are less informative than the points that are harder to predict. In our experiments on *synthetic data* ($n = 100$), this does consistently better than random subsampling as can be seen in Figure 7.



(a) RCV-1 and ℓ_1 Safe Logistic



(b) RCV-1 and ℓ_2 Squared Hinge

Figure 8: Dataset compression in classification.