Convexification of the risk

Clément Lalanne

What this lecture does

State the obstruction. The 0-1 loss is piecewise constant, so no descent method moves and exact minimization is NP-hard over a linear class.

Reduce it to one scalar function. Writing the classifier as the sign of a real valued g turns the risk into the expectation of a fixed function of the margin $yg(x)$.

Replace that function, and pay for it. A convex surrogate can be minimized. The calibration theorem says exactly what the replacement costs, and we prove it.

Compute the price once. For the logistic loss the transfer is $\epsilon \mapsto \sqrt{2\epsilon}$, derived from scratch.

Setting and notation

Binary classification. $\mathcal{Y} = \{-1, 1\}$ and $\mathbb{P}$ is a distribution on $\mathcal{X} \times \mathcal{Y}$. Write $\eta(x) = \mathbb{P}(y = 1 \mid x)$ for the **regression function**.

Risks. For a classifier $f : \mathcal{X} \to \{-1, 1\}$, $R(f) = \mathbb{P}(f(x) \neq y)$, and $R^\star = \inf_f R(f)$ over all measurable f is the Bayes risk.

Surrogate risks. For $\Phi : \mathbb{R} \to [0, \infty)$ and $g : \mathcal{X} \to \mathbb{R}$, $R_\Phi(g) = \mathbb{E}_{(x,y)}[\Phi(yg(x))]$ and $R_\Phi^\star = \inf_g R_\Phi(g)$ over all measurable g.

The question. Both infima are over everything measurable, so nothing here is about samples or estimation. The question is whether making $R_\Phi(g) - R_\Phi^\star$ small forces $R(\operatorname{sgn} \circ g) - R^\star$ to be small.

Part I. From the 0-1 loss to a margin

The 0-1 loss is combinatorial

In binary classification, $\mathcal{Y} = \{-1, 1\}$, the main loss we care about is $\ell(y, \hat{y}) = \mathbf{1}\{y \neq \hat{y}\}$. The empirical risk is then piecewise constant in the parameter.

Two obstructions, already met in the introductory lecture: the gradient vanishes almost everywhere, so no descent method moves, and exact minimization over a linear class is NP-hard in general.

The remedy, in two steps. First reparametrise the classifier as the sign of a real valued function, which turns the risk into the expectation of a function of one real variable. Then replace that function, which is discontinuous, by a convex one.

Reparametrisation as a sign test

Write $f : \mathcal{X} \to \{-1, 1\}$ as $f(x) = \operatorname{sgn}(g(x))$ for a real valued $g : \mathcal{X} \to \mathbb{R}$, with

$$\operatorname{sgn}(u) = \begin{cases} 1 & u > 0 \\ -1 & u < 0 \\ \operatorname{Unif}(\{-1, 1\}) & u = 0. \end{cases}$$

Remark. At $u = 0$ a function has been replaced by a randomised one, so the predictor now carries an auxiliary variable U, independent of everything else, and expectations are written $\mathbb{E}_{(x,y)\sim\mathbb{P},U}$. In practice $\mathbb{P}_x(g(x) = 0) = 0$ and the case never occurs, but the convention is what makes the next computation an identity rather than an inequality.

The margin identity

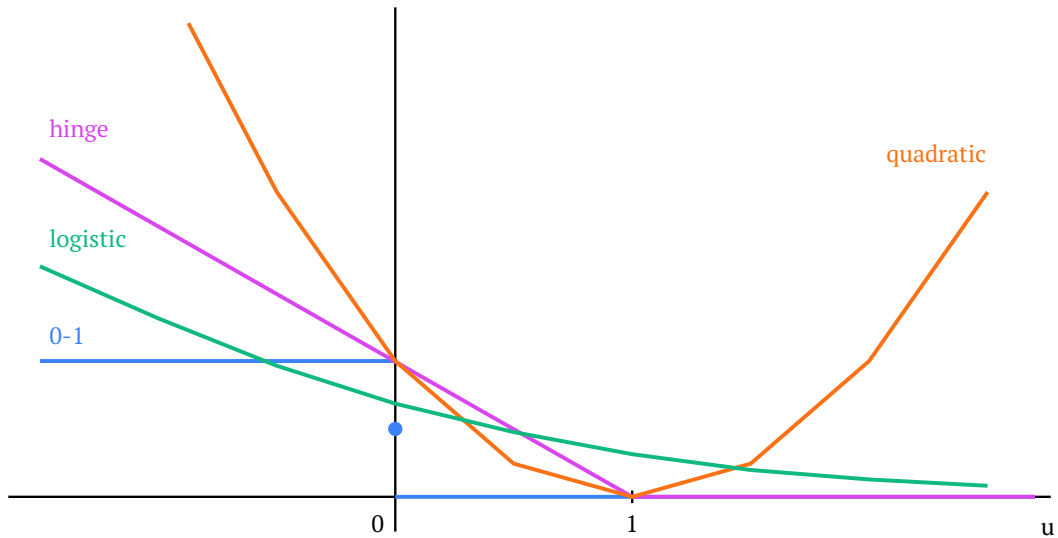
$$
\begin{aligned}
R(f) &= \mathbb{E}_{(x,y),U}[\mathbf{1}\{g(x) \neq 0\}\mathbf{1}\{f(x) \neq y\}] \\
&\qquad + \mathbb{E}_{(x,y),U}[\mathbf{1}\{g(x) = 0\}\mathbf{1}\{f(x) \neq y\}] \\
&= \mathbb{P}_{(x,y)}(yg(x) < 0) + \tfrac{1}{2}\mathbb{P}_{(x,y)}(g(x) = 0) && \text{on } \{g \neq 0\},\ f(x) \neq y \iff yg(x) < 0 \\
&= \mathbb{E}_{(x,y)\sim\mathbb{P}}[\Phi_{0\text{-}1}(yg(x))] && \Phi_{0\text{-}1}(u) = \mathbf{1}\{u < 0\} + \tfrac{1}{2}\mathbf{1}\{u = 0\}
\end{aligned}
$$

The risk of a classifier is the expectation of a fixed scalar function of the **margin** $yg(x)$. A positive margin means a correct prediction, and its size measures confidence. Everything that follows modifies $\Phi_{0\text{-}1}$ and nothing else.

Part II. Convex surrogates

Convex surrogates

Replace $\Phi_{0\text{-}1}$ by a convex $\Phi : \mathbb{R} \to [0, \infty)$ with better numerical properties, and minimize the **surrogate risk** $R_\Phi(g) = \mathbb{E}_{(x,y)\sim\mathbb{P}}[\Phi(yg(x))]$ instead.



Hinge $\Phi(u) = (1-u)_+$, quadratic $\Phi(u) = (1-u)^2$, logistic $\Phi(u) = \log(1 + e^{-u})$.

What the surrogate costs

Domination. The hinge and quadratic losses satisfy $\Phi_{0\text{-}1} \le \Phi$ pointwise, so $R(\operatorname{sgn} \circ g) \le R_\Phi(g)$ directly. The logistic loss does not: $\log(1 + e^{-u})$ equals $\log 2 < 1$ at $u = 0$. It dominates only after rescaling, and only on a compact.

The real question. Domination bounds the classification risk by the surrogate risk, but says nothing about the *excess* risks. Minimizing R_Φ is only legitimate if driving R_Φ to its infimum drives R to $R^\star$, and that is a separate statement.

Part III. What convexification costs

The conditional point of view

Definition. For $\eta \in [0,1]$, the conditional surrogate risk of a value $\alpha \in \mathbb{R}$ is $C_\eta(\alpha) = \eta\Phi(\alpha) + (1-\eta)\Phi(-\alpha)$.

Condition on x. Since $\eta(x) = \mathbb{P}(y=1 \mid x)$, the surrogate risk of the value $\alpha = g(x)$ depends on x only through $\eta(x)$:

$$\mathbb{E}_{y|x}[\Phi(yg(x))] = \eta(x)\Phi(\alpha) + (1-\eta(x))\Phi(-\alpha) = C_{\eta(x)}(\alpha).$$

So $R_\Phi(g) = \mathbb{E}_x[C_{\eta(x)}(g(x))]$. Minimizing over all measurable g is minimizing $C_\eta(\alpha)$ in α separately at each x, so $R_\Phi^\star = \mathbb{E}_x[H(\eta(x))]$ with $H(\eta) = \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha)$.

The same for $\Phi_{0\text{-}1}$. Its conditional risk equals $1-\eta$ for $\alpha > 0$, η for $\alpha < 0$, and $\frac{1}{2}$ at $\alpha = 0$. The minimum is $\min(\eta, 1-\eta)$, attained exactly when α has the sign of $2\eta - 1$: that is the Bayes rule.

The excess risks, pointwise

Write $e_\eta(\alpha)$ for the excess classification risk at a point, the conditional 0-1 risk of α minus $\min(\eta, 1-\eta)$.

$$e_\eta(\alpha) = \begin{cases} 0 & \alpha(2\eta - 1) > 0 \\ \frac{1}{2}|2\eta - 1| & \alpha = 0 \\ |2\eta - 1| & \alpha(2\eta - 1) < 0. \end{cases}$$

In particular $e_\eta(\alpha) \le |2\eta - 1| \mathbf{1}\{\alpha(2\eta - 1) \le 0\}$: a wrong or undecided sign is the only way to pay anything.

Averaging over x,

$$R(\operatorname{sgn} \circ g) - R^\star = \mathbb{E}_x[e_{\eta(x)}(g(x))],$$
$$R_\Phi(g) - R_\Phi^\star = \mathbb{E}_x[C_{\eta(x)}(g(x)) - H(\eta(x))].$$

Both excess risks are averages of pointwise quantities, so it is enough to compare them at a fixed η. The question becomes: when α has the wrong sign, how much surrogate risk must be paid?

The calibration function

Definition. For $\eta \in [0,1]$ set

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} C_\eta(\alpha), \quad H^-(\eta) = \inf_{\alpha(2\eta-1) \le 0} C_\eta(\alpha),$$

the second being the best a predictor with the wrong sign can do. For $\theta \in [0,1]$ put

$$\tilde{\psi}(\theta) = H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right),$$

and let ψ be the largest convex function on $[0,1]$ below $\tilde{\psi}$.

Three properties. $\tilde{\psi} \ge 0$, since the second infimum runs over a smaller set. $\tilde{\psi}(0) = 0$, since at $\eta = 1/2$ the constraint is vacuous. And ψ is nondecreasing: for $0 \le a \le b$, convexity and $\psi(0) = 0$ give $\psi(a) \le \frac{a}{b}\psi(b) \le \psi(b)$.

The pointwise comparison

Lemma. For every $\eta \in [0,1]$ and every $\alpha \in \mathbb{R}$,

$$\psi(e_\eta(\alpha)) \le C_\eta(\alpha) - H(\eta).$$

Proof. Put $\theta = |2\eta - 1|$, so that $\tilde{\psi}(\theta) = H^-(\eta) - H(\eta)$.

If $\alpha(2\eta - 1) > 0$ then $e_\eta(\alpha) = 0$, the left side is $\psi(0) = 0$, and the right side is nonnegative.

Otherwise α is admissible in the infimum defining H^-, so $C_\eta(\alpha) \ge H^-(\eta)$ and

$$C_\eta(\alpha) - H(\eta) \ge H^-(\eta) - H(\eta) = \tilde{\psi}(\theta) \ge \psi(\theta) \ge \psi(e_\eta(\alpha)),$$

using $\psi \le \tilde{\psi}$, then $e_\eta(\alpha) \le \theta$ with ψ nondecreasing. ∎

The calibration theorem

Theorem. For every measurable $g : \mathcal{X} \to \mathbb{R}$,

$$\psi\left(R(\operatorname{sgn} \circ g) - R^\star\right) \le R_\Phi(g) - R_\Phi^\star.$$

Proof. One line per tool.

$$\begin{aligned}
\psi(R(\operatorname{sgn} \circ g) - R^\star) &= \psi\left(\mathbb{E}_x[e_{\eta(x)}(g(x))]\right) && \text{pointwise form of the excess risk} \\
&\le \mathbb{E}_x\left[\psi(e_{\eta(x)}(g(x)))\right] && \text{Jensen, } \psi \text{ convex} \\
&\le \mathbb{E}_x\left[C_{\eta(x)}(g(x)) - H(\eta(x))\right] && \text{the lemma, at each } x \\
&= R_\Phi(g) - R_\Phi^\star
\end{aligned}$$

■

What the theorem says

Consistency. If ψ is continuous and vanishes only at 0, then $R_\Phi(g_m) \to R_\Phi^\star$ forces $R(\operatorname{sgn} \circ g_m) \to R^\star$. Minimizing the surrogate is then a legitimate substitute, not a different problem.

The rate is the shape of ψ at the origin. A ψ that is linear near 0 transfers excess risk one for one. A ψ that is quadratic near 0 costs a square root, which is the generic case for smooth losses.

Nothing here involves a sample. This is a statement about two population risks. The cost of convexification and the cost of estimation are separate, and they add.

Part IV. The logistic loss, in full

Its conditional risk

Take $\Phi(u) = \log(1 + e^{-u})$, so that $C_\eta(\alpha) = \eta \log(1 + e^{-\alpha}) + (1 - \eta) \log(1 + e^{\alpha})$ and

$$C'_\eta(\alpha) = -\frac{\eta}{1 + e^{\alpha}} + \frac{(1 - \eta)e^{\alpha}}{1 + e^{\alpha}} = \frac{(1 - \eta)e^{\alpha} - \eta}{1 + e^{\alpha}},$$

which vanishes exactly at the log odds $\alpha^\star(\eta) = \log \frac{\eta}{1-\eta}$.

Writing $\sigma(u) = 1/(1 + e^{-u})$, we have $\sigma(\alpha^\star) = \eta$, hence $\log(1 + e^{-\alpha^\star}) = -\log \eta$ and $\log(1 + e^{\alpha^\star}) = -\log(1 - \eta)$, so

$$H(\eta) = -\eta \log \eta - (1 - \eta) \log(1 - \eta) =: H_b(\eta).$$

Minimizing the logistic risk recovers the conditional probability, and the value at the optimum is the entropy left in y once x is known.

Its calibration function

For $\eta > 1/2$ the constraint in H^- reads $\alpha \leq 0$, while C_η is convex with its unconstrained minimum at $\alpha^\star > 0$. The constrained minimum therefore sits at the boundary:

$$H^-(\eta) = C_\eta(0) = \eta \log 2 + (1-\eta)\log 2 = \log 2.$$

Hence

$$\tilde{\psi}(\theta) = \log 2 - H_b\left(\frac{1+\theta}{2}\right),$$

the relative entropy between the Bernoulli law of parameter $\frac{1+\theta}{2}$ and the fair coin.

No convexification is needed, so $\psi = \tilde{\psi}$: differentiating,

$$\psi'(\theta) = \frac{1}{2}\log\frac{1+\theta}{1-\theta}, \qquad \psi''(\theta) = \frac{1}{1-\theta^2} > 0.$$

The quadratic lower bound

Claim. $\psi(\theta) \geq \theta^2/2$ for $\theta \in [0, 1)$.

Proof. Let $h(\theta) = \psi(\theta) - \theta^2/2$. Then

$$h(0) = \log 2 - H_b(\tfrac{1}{2}) = 0,$$
$$h'(0) = \psi'(0) = 0,$$
$$h''(\theta) = \frac{1}{1-\theta^2} - 1 = \frac{\theta^2}{1-\theta^2} \geq 0.$$

So h is convex with a stationary point at the origin, hence minimal there, and $h \geq h(0) = 0$. ∎

This is Pinsker's inequality for two Bernoulli laws, proved here in the one case we need.

The bound for the logistic loss

Corollary. For every measurable g,

$$R(\operatorname{sgn} \circ g) - R^\star \le \sqrt{2\,(R_\Phi(g) - R_\Phi^\star)}.$$

Proof. Write $e = R(\operatorname{sgn} \circ g) - R^\star$ and $\epsilon = R_\Phi(g) - R_\Phi^\star$. The theorem gives $\psi(e) \le \epsilon$ and the claim gives $e^2/2 \le \psi(e)$, so $e^2 \le 2\epsilon$. ∎

Reading it. An excess logistic risk ϵ buys an excess classification risk $\sqrt{2\epsilon}$, not ϵ. Halving the surrogate excess improves the guarantee only by $\sqrt{2}$. That is the price of a smooth loss, and it is paid before any sample is drawn.

Other losses, for comparison

In each case $\alpha = 0$ is admissible and C_η is convex with $\alpha^\star \geq 0$ for $\eta \geq 1/2$, so $H^-(\eta) = \Phi(0)$ and only H has to be computed.

$\Phi(u)$	$H(\eta)$	$\psi(\theta)$
hinge, $(1-u)_+$	$2\min(\eta, 1-\eta)$	θ
quadratic, $(1-u)^2$	$4\eta(1-\eta)$	θ^2
exponential, e^{-u}	$2\sqrt{\eta(1-\eta)}$	$1 - \sqrt{1-\theta^2}$
logistic, $\log(1+e^{-u})$	$H_b(\eta)$	$\log 2 - H_b\left(\frac{1+\theta}{2}\right)$

The hinge loss transfers excess risk one for one, which is the best possible. The other three are quadratic at the origin, hence a square root: θ^2 for the quadratic loss and $\theta^2/2$ to second order for the other two.

What to remember

The margin identity. $R(\operatorname{sgn} \circ g) = \mathbb{E}[\Phi_{0\text{-}1}(yg(x))]$. The whole problem is one scalar function of the margin, and convexification changes that function and nothing else.

Calibration is pointwise. Both excess risks are averages over x of quantities depending only on $\eta(x)$ and $g(x)$, so the comparison is a question about one number.

The calibration function. ψ is the convexified gap between the best wrong-signed conditional risk and the best conditional risk. The theorem is the pointwise lemma plus Jensen.

The price, for the logistic loss. $R(\operatorname{sgn} \circ g) - R^\star \le \sqrt{2(R_\Phi(g) - R_\Phi^\star)}$, from $\psi(\theta) = \log 2 - H_b(\frac{1+\theta}{2}) \ge \theta^2/2$.