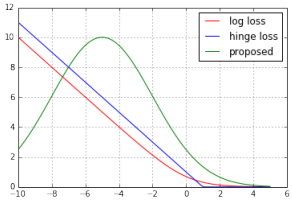
Theoretical homework 2
due April 27 (Wednesday), 23:59.

Innopolis
Machine learning course

Recommendations: all solutions should be short, mathematically strict (unless qualitative explanation is needed), precise with respect to the stated question and clearly written. The scan of the solution should be sent to v.v.kitov@yandex.ru with your name, surname and word "innopolis" in the title. Please send before deadline, because late submissions will be downgraded by 50%.

1. **Non-monotonous loss:** Loss functions that we considered for linear classifier weight estimation were monotonously decreasing, such as log-loss $\mathcal{L}(M) = \ln(1 + e^{-M})$ or hinge loss $\mathcal{L}(M) = [1 - M]_+$. On picture below you are proposed a non-monotonous loss function. Please give an idea in what cases this loss function may be useful.



2. **Gradient optimization:** For Perceptron of Rosenblatt method $\mathbb{I}[M < 0] \approx \mathcal{L}(M) = [-M]_+ = max\{-M, 0\}$ and for logistic regression $\mathbb{I}[M < 0] \approx \mathcal{L}(M) = \ln(1 + e^{-M})$. For both methods:

   (a) Plot $\mathcal{L}(M)$ on the same graph

   (b) Plot $\frac{\partial \mathcal{L}(M)}{\partial M}$ on the same graph

   (c) Write down the update rule of weights for stochastic gradient descent method

   (d) Looking at the results of a), b) and c), what is the qualitative difference between the two methods?

   (e) Write down the update rule of weights for gradient descent method.

   (f) What is the advantage of c) compared to e) update formula?

3. **Regularized regression:** Consider D-dimensional feature space $x \in \mathbb{R}^D$, design matrix $X \in \mathbb{R}^{N \text{x} D}$. We have derived estimate of β in linear regression $\widehat{y} = \beta^T x$, β, using least squares estimation:

$$\sum_{n=1}^{N} (\beta^T x_n - y_n)^2 \to \min_{\beta}$$

   (a) will that solution exist when $X^T X$ is degenerate matrix?

   (b) Derive estimate of β when we use least-squares minimization with regularization:

$$\sum_{n=1}^{N} (\beta^T x_n - y_n)^2 + \sum_{d=1}^{D} \beta_d^2 \to \min_{\beta}$$

   (c) will the solution for case (b) exist when $X^T X$ is degenerate matrix?