# Multi-class Classification: Mirror descent approach

Daria Reshetova
EE department, Stanford University

## Problem overview

Common multi-class classification approaches:

- Naive Bayes (text classification)
- Softmax (image classification, NN)
- one-vs-all, all-vs-all, ECOC + SVM (small $k$)

Practical setup:

- large number of classes $k$
- high input space dimension $n$
- limited number of labeled instances $m$

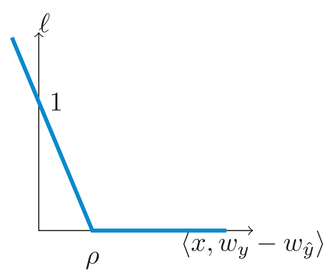
## Goals

- provide a way to extend SVMs to multiclass classification in a scalable manner
- evaluate the dependance of the generalization error on the number of classes
- utilize problem's structure for better performance

## Optimization problem

$$\ell(x,y) = \max_{\hat{y}\neq y}\{0, 1 - \langle x, w_y - w_{\hat{y}}\rangle / \rho\}$$

$$\mathbb{E}\,\ell(x,y) \to \min_{w\in\mathbb{R}^{n\times k}}$$




The margins in this case are $m(x,y,w) = \langle x, w_y - w_{\hat{y}}\rangle$. They represent the difference between the true class closest wrong class $\Longrightarrow$ similar to 1-vs-all classification with SVM

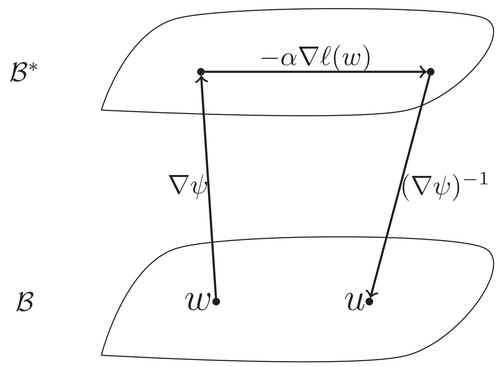
Classification is performed as: $y = \arg\max_y r\langle x, w_y\rangle$

## Algorithm

**Stochastic descent step:**

$$w^{i+1} = \arg\min_w \{\alpha_i \langle \nabla_w \ell(x_i, y_i, w), w\rangle + \overset{\Delta(w,w^i)}{\|w, w^i\|^2}\}$$

The $\ell_2-$ distance here is substituted by $\Delta(w,u) = \psi(w) - \psi(u) - \langle \psi'(u), w-u\rangle$, where $\psi(w)$ is some strongly convex function. The idea is that $\psi(w)$ maps to the dual space, allowing norms other than $\ell_2$ to measure gradient.




The freedom in choosing $\psi$ results in a variety of algorithms. The main reason for doing that is the difference in distance measurement between the initial point and the optimal solution, that depends on $k$. Let $\Omega^2 = \Delta(w^0, w^*)$

**Common settings**

1. $\|w_i^0 - w_i^*\|_2 \le 1$ and $\psi(w) = \sum_{i=1}^n \|w_i\|^2/2$ implies Euclidian setup and $\Omega = O(\sqrt{k})$
2. $\sum_{i=1}^k \|w_i^0 - w_i^*\|_2 \le 1$ ($\ell_1/\ell_2$ ball)
   $\psi(w) = 2e\ln k \sum_{i=1}^k \|w_i\|^{1+1/\ln 2k}$ implies $\ell_1/\ell_2$ norm and $\Omega^2 = O(\ln k)$
3. $\|w_i^0 - w_i^*\|_2 \le 1\,\forall i\}$ and
   $\psi(w) = 2e\ln k \sum_{i=1}^k \|w_i\|^{1+1/\ln 2k}$ implies $\Omega = O(k\ln k)$

## Dataset

**ALOI dataset**

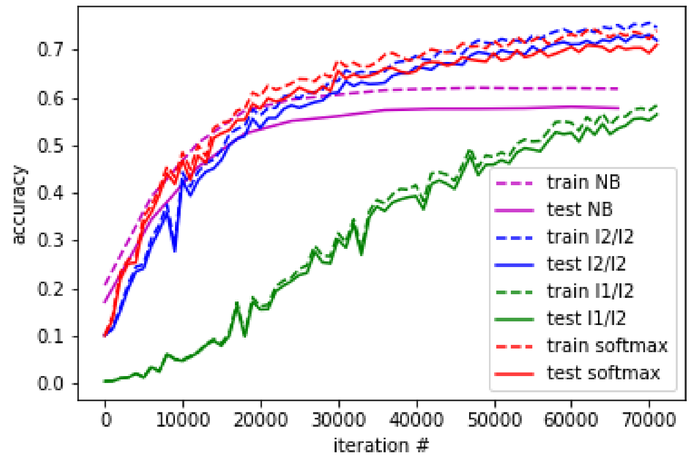
- color images
- classes: 1000
- 108 images / class

## Algorithms tested

The results were compared to common multiclass classification algorithms:

- Naive Bayes classifier
- linear regression
- $\ell_1/\ell_2, \ell_2/\ell_2$ setup balls

## Experimental results




## Theoretical results

**General case**
For $G = \sqrt{\sup_{w\in\mathcal{W}} \mathbb{E}\,\|g\|^2}$ and constant steps $\alpha_m = \frac{\Omega}{G\sqrt{n}}$ the excess risk rate is

$$\mathbb{E}\left[\ell(x,y,w^{(n)}) - \ell(x,y,w^*)\right] \le \frac{\Omega G}{2\sqrt{n}}$$

**Common settings rates**:

1. $O(\sqrt{k/n})$ rate for the Euclidian setup
2. $O(\ln k/\sqrt{n})$ rate for $\ell_1/\ell_2$ ball
3. $O(k\ln k/\sqrt{n})$ for setting 3

Rates 1,2 are asymptotically unimprovable in general case.