# Multi-class classification via proximal mirror descent

Daria Reshetova
Stanford
EE department
resh@stanford.edu

## Abstract

*We consider the problem of multi-class classification and a stochastic optimization approach to it. The idea is to, instead of weighing classes, make use of the total sum of margins as a regularization. As general the problem is hard to solve, we use Bregman divergence as the regularizer and end up with a proximal mirror descent with a specific distance-generating function. The approach is designed for problems with highly unbalanced classes as it makes use of different margins between each class and the rest, therefore emulating the one-vs-all approach. This should decrease the error dependence in terms of the number of classes.*

## 1. Introduction (Motivation)

Classification is one of the core machine learning tasks. Multi-class classification arises in various problems including document classification [16], image [6, 10], gesture [11] and video recognition [9] and many others. Datasets for the problems are growing in both number of samples and number of classes $k$. As the expected error of classification algorithms also increases, the growth rate in terms of $k$ becomes crucial.

In this project we consider the classical one-vs-all margin classification approach [1], which was empirically shown to be as good as ECOC and all-vs-all approach, at least from the practical point of view [17]. There are several generalization ability guarantees known for the class of learners. Distribution-independent ones rely on function class complexity measures such as Natarajan [13], graph [13, 5] and Vapnic-Chervoninkis [7] dimension. If we consider kernel separators $f_y(x, w) = K(x, w_y)$ with PSD K, the bounds lead to $\tilde{O}(k/\sqrt{n})$ excess risk [4, 7], Covering number based bound presented in [20] gives $\tilde{O}(\sqrt{k/n})$ rate. While the bounds provided seem tight, they result in large constants and dimension dependence for combinatorial complexity measures. As for distribution-dependent bounds, little is known beyond the general Rademacher complexity based bound, which is in the worst case of order $O(k/\sqrt{n})$ for typical function classes (e. g. finite VC-dimensional).

As the underlying distribution of the pairs object-class is unknown, the problem can be solved by substituting the actual risk by the empirical one or by means of stochastic optimization. While the general problem statement of minimizing the risk allows the implementation of different optimization algorithms, first order methods are preferable to high-order ones for large-scale problems in terms of their generalization ability and computational efficiency [3]. We consider an adaptation of stochastic Mirror Descent algorithm [15, 2] to solve the problem. Mirror Descent is a first-order algorithm for convex function minimization, which restricts the method to convex $\Omega$ and convex in $w$ loss-function $\ell(x, y, w)$, allowing dimension-independent excess risk bounds.

## 2. Method

The method we propose would be proximal mirror descent with various distance generating functions to regularize the standard classifier. The idea of the method is to use mirror descent to focus on minimizing the expected loss rather than the usual empirical risk to increase the model's generalization.

### 2.1. Framework

We first describe the framework of multi-class classification used in the paper. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a set of instances, $\mathcal{Y} = \{1 \dots k\}$ be a set of classes. The general assumption is that $\mathcal{X} \times \mathcal{Y}$ support a probability space $(\mathcal{X} \times \mathcal{Y}, A, \mathbb{P})$ and the sample $S = \{x_i, y_i\}_{i=1}^n$ is i.i.d. drawn from the distribution. We denote $\mathcal{F} = \{f(\cdot, \cdot, w) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} | w \in \mathcal{W}\}$ the class of decision functions – a parametric class of measurable functions and $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \to \mathbb{R}_+$ the loss function. We consider one-vs-all approach to the problem by setting the predictor to be $\hat{f}(x, w^*) = \max_{y \in \mathcal{Y}} f(x, y, w^*)$. The loss function is chosen to be a Lipschitz upper bound on the indicator function $[\hat{f}(x_i, w) = y_i]$ and to maximize the mar-

gin:

$$m(x, y, w) = f(x, y, w) - \max_{\hat{y} \in \mathcal{Y},\, \hat{y} \neq y} f(x, \hat{y}, w)$$

We use $\ell(x, y, w) = \max\{0, 1 - m(x, y, w)/\rho\}$ as the loss finction. The problem is then to minimize the expected risk $F(w) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\ell(x, y, w)$ :

$$w^* = \arg\min_{w \in \mathcal{W}} \mathbb{E}_{(x,y)}\ell(m(x, y, w)) \tag{1}$$

To make the application of mirror descent possible, the underlying function has to be convex, the fact that the distribution $\mathbb{P}$ is unknown leads to $f(x, y, w)$ needing to be linear in $w$, so $\mathcal{F} = \{\langle x, w_y\rangle \mid w \in \mathcal{W}\}$ further in the paper. The case can also be generalized to PSD-kernel classification via linear classifiers in RKHS [12].

## 2.2. Algorithm

Mirror descent algorithm is similar to stochastic gradient descent, except that for $\mathcal{W} \subset E$ with $E$ being a Euclidean space it ensures gradient steps to be made in $E^*$ by mapping there with $\nabla\psi$, where $\psi : E \to \mathbb{R}$ is a strongly convex function with gradient field continuous on $\mathcal{W}$:

$$\psi(w^1) - \psi(w^2) - \left\langle \nabla\psi(w^2), w^1 - w^2 \right\rangle \geq \frac{1}{2}\|w^1 - w^2\|^2.$$

As the number of classes is a factor of the dimension of $\mathcal{W}$, choosing the right proximity to measure the set diameter can effectively lower the error rate. where $\Delta(w^1, w^2) =$

**Data:** pairs $(x^{(i)}, y^{(i)})_{i=1}^m$, stepsizes $(\alpha_i)_{i=1}^m$
**Result:** $w^{(m)}$ the matrix of separating hyperplanes
initialization;
i = 0;
**while** $i < m$ **do**
  $w^{i+1};=$
  $\arg\min_{w \in W} \{\Delta(w, w^i) + \alpha_m\langle F'(w^i), w - w^i\rangle\},;$
  i++;
**end**
$w^{(m)} := \sum_{i=1}^m \alpha_i w^i / \sum_{i=1}^m \alpha_i;$
**Algorithm 1:** Mirror descent

$\psi(w^1) - \psi(w^2) - \left\langle \nabla\psi(w^2), w^1 - w^2 \right\rangle \geq \frac{1}{2}\|w^1 - w^2\|^2$. In case of expectation minimization gradients are taken at random points $g^k \in \partial\ell(x_k, y_k, w^k)$, which ensures $\mathbb{E}g^k \in \partial F(w^k)$ as long as $(x_k, y_k)$ and $w^k$ are independent.

## 2.3. Theoretical results

To measure the generalization ability of the algorithm we use the expectation of the difference between the loss function taken at a point $w^{(n)}$ and the optimal point.

### 2.3.1 Oracle inequalities

Mirror descent algorithm is similar to stochastic gradient descent, except that for $\mathcal{W} \subset E$ with $E$ being a Euclidean space it ensures gradient steps to be made in $E^*$ by mapping there with $\nabla\psi$, where $\psi : E \to \mathbb{R}$ is a strongly convex function with gradient field continuous on $\mathcal{W}$:

$$\psi(w^1) - \psi(w^2) - \left\langle \nabla\psi(w^2), w^1 - w^2 \right\rangle \geq \frac{1}{2}\|w^1 - w^2\|^2.$$

As the number of classes is a factor of the dimension of $\mathcal{W}$, choosing the right proximity to measure the set diameter can effectively lower the error rate.

Mirror Descent steps are gradient steps with Bregman divergence of $\psi$ in the role of the distance:

$$w^1 = \arg\min_{w \in \mathcal{W}} \psi(w)$$

$$w^{m+1} = \arg\min_{w \in W} \{\Delta(w, w^m) + \alpha_m\langle F'(w^m), w - w^m\rangle\},$$

$\Delta(w^1, w^2) = \psi(w^1) - \psi(w^2) - \left\langle \nabla\psi(w^2), w^1 - w^2 \right\rangle \geq \frac{1}{2}\|w^1 - w^2\|^2$. In case of expectation minimization gradients are taken at random points $g^k \in \partial\ell(x_k, y_k, w^k)$, which ensures $\mathbb{E}g^k \in \partial F(w^k)$ as long as $(x_k, y_k)$ and $w^k$ are independent.

**Lemma 1.** *[14, 2] For all $w \in \mathcal{W}$*

$$\begin{aligned} \Delta(w, w^{m+1}) &\leq \alpha_m \left\langle g^m, w - w^{m+1} \right\rangle \\ &+ \quad +\Delta(w, w^m) - \Delta(w^{m+1}, w^m) \end{aligned}$$

*Proof.* Set $h(w) = \Delta(w, w^m) + \alpha_m\langle g^m, w - w^m\rangle$, then $w^{m+1} = \arg\min_{w \in \mathcal{W}} h(w)$.
Optimality of $w^{m+1}$ leads to $\left\langle h'(w^{m+1}), w - w^{m+1} \right\rangle \geq 0$

As long as $h'(w^{m+1}) = \nabla\psi(w^{m+1}) - \nabla\psi(w^m) + \alpha_m g^m$, we can rearrange the terms and get

$$\begin{aligned} 0 &\leq \left\langle \alpha_m g^m + \nabla\psi(w^{m+1}) - \nabla\psi(w^m), w - w^{m+1} \right\rangle \\ &= \left\langle \alpha_m g^m, w - w^{m+1} \right\rangle - \langle \nabla\psi(w^m), w - w^m \rangle \\ &+ \left\langle \nabla\psi(w^{m+1}), w - w^{m+1} \right\rangle \\ &+ \left\langle \nabla\psi(w^m), w^{m+1} - w^m \right\rangle \\ &= \left\langle \alpha_m g^m, w - w^{m+1} \right\rangle - \Delta(w, w^{m+1}) \\ &+ \Delta(w, w^m) - \Delta(w^{m+1}, w^m) \end{aligned}$$

□

Lemma 1 and the fact that $\mathbb{E}g_k \in \partial F(w^k)$ result in an oracle inequality for stochastic Mirror Descent.

**Corollary 1.** *For $U^2 = \arg\max_{u,w \in \mathcal{W}} (\psi(u) - \psi(w))$, $G^2 = \arg\max_{w \in \mathcal{W}} \mathbb{E}\|g(x, y, w)\|^2$, with $g(x, y, w) \in \partial\ell(x, y, w)$ and*

*for any* $w \in \mathcal{W}$ *and* $w^{(n)} = \frac{\sum_{m=1}^{n} \alpha_m w^m}{\sum_{m=0}^{n} \alpha_m}$

$$\mathbb{E}\left(\ell(x,y,w^{(n)}) - \ell(x,y,w)\right) \leq \frac{U^2 + G^2 \sum_{m=1}^{n} \alpha_m^2/2}{\sum_{m=1}^{n} \alpha_m} \tag{2}$$

*Proof.* According to 1 and by the strong convexity of $\psi(w)$ :

$$\begin{aligned}
\Delta(w, w^{m+1}) &\leq \alpha_m \left\langle g^m, w - w^{m+1} \right\rangle + \Delta(w, w^m) \\
&- \Delta(w^{m+1}, w^m) \\
&\leq \langle \alpha_m g^m, w - w^m \rangle \\
&+ \left\langle \alpha_m g^m, w^m - w^{m+1} \right\rangle + \Delta(w, w^m) - \\
&- \frac{1}{2}\|w^{m+1} - w^m\|^2 \\
&\leq \langle \alpha_m g^m, w - w^m \rangle \\
&+ \alpha_m \|g^m\| \|w^m - w^{m+1}\| + \Delta(w, w^m) - \\
&- \frac{1}{2}\|w^{m+1} - w^m\|^2
\end{aligned}$$

Summing over $m = 1, \ldots, n$ gives:

$$0 \leq \sum_{m=0}^{n} \langle \alpha_m g^m, w - w^m \rangle + \frac{1}{2} \sum_{m=1}^{n} \alpha_m^2 \|g_m\|^2 + \Delta(w, w^1) \tag{3}$$

As $w^1 = \arg\min_{w \in \mathcal{W}} \psi(w)$ :

$$\begin{aligned}
\Delta(w, w^1) &= \psi(w) - \psi(w^1) - \left\langle \psi'(w^1), w^1 - w \right\rangle \\
&\leq \psi(w) - \psi(w^1) \leq U^2,
\end{aligned}$$

By convexity of $\ell$ and the independence of $(x_m, y_m)$ and $w^m$ :

$$\begin{aligned}
&\sum_{m=0}^{n} \alpha_m (\ell(x_m, y_m, w^m) - \ell(x_m, y_m, w)) \\
&\qquad \leq \frac{1}{2} \sum_{m=1}^{n} \alpha_m^2 \|g_m\|^2 + U^2, \\
&\sum_{m=1}^{n} \alpha_m \left(F(w^{(n)}) - F(w)\right) \\
&\leq \sum_{m=1}^{n} \alpha_m \left(\mathbb{E}\left(\ell(x,y,w^m) - \ell(x,y,w)\right)\right) \\
&\qquad \leq \frac{1}{2} \sum_{m=1}^{n} \alpha_m^2 G^2 + U^2
\end{aligned}$$

□

If steps are constant $\alpha_m = \alpha = \frac{\sqrt{2}U}{G\sqrt{n}}$ we get

$$F(w^{(n)}) - F(w^*) \leq \frac{\sqrt{2}UG}{\sqrt{n}}$$

The dependence in the number of classes in the excess risk bound is hidden in constants $U$ and $G$ and depends on the choice of distance-generating function $\psi(w)$ and the sets $\mathcal{W}, \mathcal{X}$.

For $U^2 = \arg\max_{u,w \in \mathcal{W}} \left(\psi(u) - \psi(w)\right)$, $G^2 = \arg\max_{w \in \mathcal{W}} \mathbb{E}\|g(x,y,w)\|^2$, with $g(x,y,w) \in \partial \ell(x,y,w)$ and for any $w \in \mathcal{W}$ and $w^{(n)} = \frac{\sum_{m=1}^{n} \alpha_m w^m}{\sum_{m=0}^{n} \alpha_m}$

$$\mathbb{E}\left(\ell(x,y,w^{(n)}) - \ell(x,y,w)\right) \leq \frac{U^2 + G^2 \sum_{m=1}^{n} \alpha_m^2/2}{\sum_{m=1}^{n} \alpha_m} \tag{4}$$

### 2.4. Proximal setups

1. Consider $\mathcal{W} = \{w \in \mathbb{R}^{d \times k} | \max_i \|w_i\|_2 \leq \Omega\}$, $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_2 < X\}$ and $\psi(w) = \frac{1}{2}\sum_{i=1}^{k} \|w_i\|_2^2$. In this case $\Delta(w^1, w^2) = \frac{1}{2}\sum_{i=1}^{k} \|w_i^1 - w_i^2\|_2^2$ and $U^2 = k\Omega^2$, $G^2 = \frac{2X^2}{\rho^2}$. This leads to excess risk rate

$$F(w^{(n)}) - F(w^*) \leq \frac{2\Omega X}{\rho}\sqrt{\frac{k}{n}}$$

2. If the margins are allowed to be different for different classes: $\mathcal{W} = \{w \in \mathbb{R}^{d \times k} | \sum_{i=1}^{k} \|w_i\|_2 \leq \Omega\}$ and $\psi(w)$ is chosen to be strongly convex w.r.t. $\ell_1/\ell_2$ norm: $\psi(w) = \frac{e \ln k}{1 + 1/\ln k} \sum_{i=1}^{k} \|w^i\|_2^{1+1/\ln k}$ [8], which is the case of favorable geometry, then $U^2 = e \ln k \Omega$, $G = X/\rho$ and the rate can be pushed to

$$F(w^{(n)}) - F(w^*) \leq \frac{X}{\rho} \frac{\sqrt{2e\Omega \ln(k)}}{\sqrt{n}}$$

3. We have also tested various other norm regularizations:

$$\psi(w) = O(1) \left(\sum_{i=1}^{k} \|w_i\|_2^p\right)^{1/p},$$

which corresponds to the $\ell_p$-norm of the margin's Euclidian norm. If the initial point was chosen so that $\max_i \|w_i\| \leq \Omega$, then the excess risk is of order:

$$F(w^{(n)}) - F(w^*) \leq \frac{X}{\rho} \frac{\Omega \sqrt[p]{k}}{\sqrt{n}}$$

## 3. Dataset

The dataset used in the project is the aloi dataset:[18] a color image collection of one-thousand small objects, recorded for scientific purposes. In order to capture the sensory variation in object recordings, the viewing angle, illumination angle, and illumination color for each object

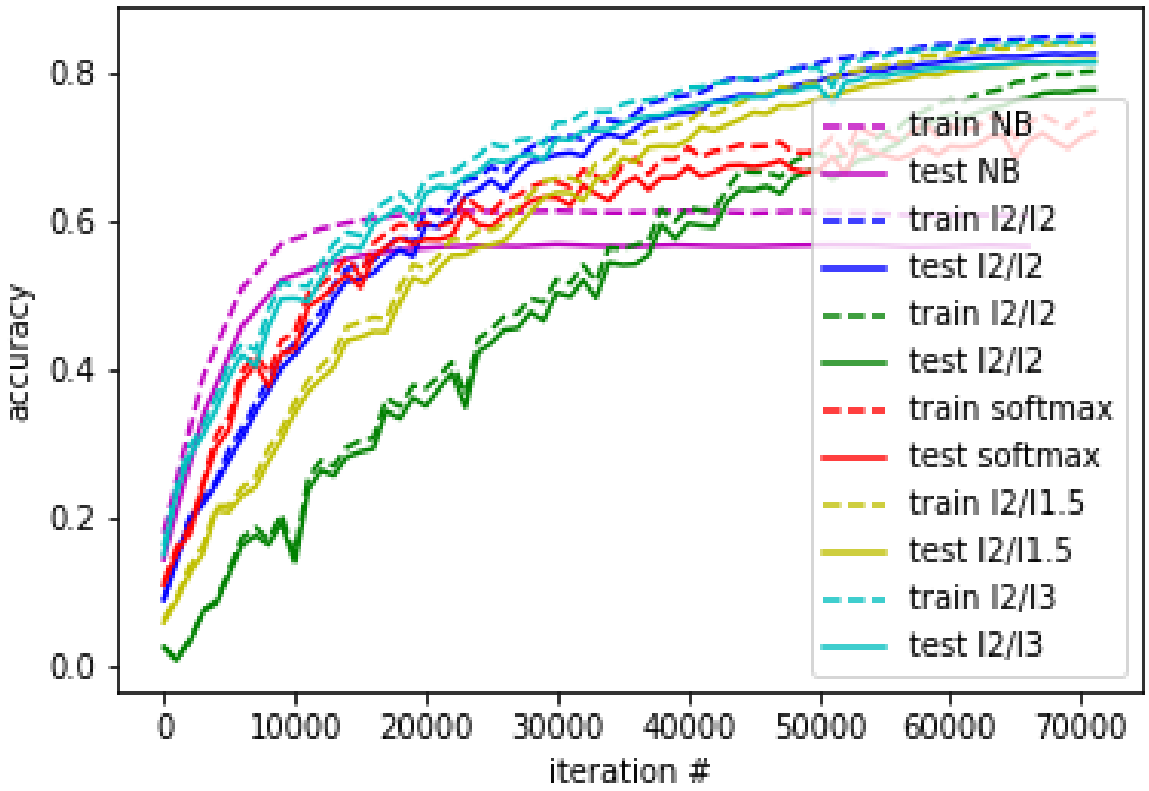


Figure 1: Experiment result

were systematically varied, and additionally wide-baseline stereo images were captured (the description taken from http://aloi.science.uva.nl). The dataset consists of 108000 images with 1000 classes and 108 images per class. This is an example of a well-balanced dataset.

## 4. Experiments

To evaluate the performance of mirror descent on the dataset we have compared it to the Naive Bayes and the softmax-loss linear classifier. We haven't used svm as it took too long to run. We have also tried stochastic versions of svm like Pegasos[19], but training on the dataset in a binary way in general was not a good idea as the dataset for one-vs-all approach was very unbalanced. The $\ell_2/\ell_2$ setup for the mirror descent algorithm can be viewed as the one-vs-all svm, as the regularization used is exactly the same and the loss is also the same, but the major difference with pegasos is that the optimization is performed simultaneously for all classes.

Data preprocessing was to scale the features to have 0-meam and 1-variance. The norms tested were $\ell_2/\ell_p$ norms with $p \in \{1, 1.5, 2, 3\}$. The results are presented on the plot above. The final accuracies were:

| Algorithm | Accuracy, % |
|---|---|
| $\ell_2/\ell_1$ norm | 78.2 |
| $\ell_2/\ell_{1.5}$ norm | 82.1 |
| $\ell_2/\ell_2$ norm | 82.8 |
| $\ell_2/\ell_3$ norm | 82.5 |
| Naive Bayes | 56.6 |
| SoftMax classifier | 72.1 |

According to the experiments, $\ell_2/\ell_2$ and $\ell_2/\ell_3$ norms provided best results, despite the theoretical guarantees. There might be several reasons for that. First of all, the plot shows that the curve for $\ell_2/\ell_1$ setup still has potential for growth, so the dataset size can be too small for the weights to reach the desired accuracy. Second, the dataset was too balanced for the problem.

## 5. Future work

My plans for the future work include testing the algorithm on more datasets, especially the unbalanced ones to

see how it behaves and to come up with a smarter way of choosing the stepsizes rather than doing gridsearch.

## References


[1] Mohamed Aly. Survey on multiclass classification methods. *Neural Netw*, pages 1–9, 2005.

[2] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[3] Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

[4] Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *arXiv preprint arXiv:1308.2893*, 2013.

[5] Richard M Dudley. Universal donsker classes and metric entropy. In *Selected Works of RM Dudley*, pages 345–365. Springer, 2010.

[6] Giles M Foody and Ajay Mathur. A relative evaluation of multiclass image classification by support vector machines. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(6):1335–1343, 2004.

[7] Yann Guermeur. Vc theory of large margin multi-category classifiers. *The Journal of Machine Learning Research*, 8:2551–2594, 2007.

[8] Anatoli Juditsky and Arkadi Nemirovski. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011.

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[10] Daniel D Lee and H Sebastian Seung. Unsupervised learning by convex and conic coding. *Advances in neural information processing systems*, pages 515–521, 1997.

[11] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*, chapter 8, pages 198–200. MIT press, 2012.

[13] Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[14] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[15] A-S Nemirovsky, D-B Yudin, and E-R Dawson. Problem complexity and method efficiency in optimization. 1982.

[16] Jason DM Rennie and Ryan Rifkin. Improving multiclass text classification with the support vector machine. 2001.

[17] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

[18] Anderson Rocha and Siome Goldenstein. Multiclass from binary: Expanding one-vs-all, one-vs-one and ecoc-based approache. *IEEE Transactions on Neural Networks and Learning Systems, 25(2):289-302*, 2014.

[19] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[20] Tong Zhang. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.