Mehryar Mohri
Advanced Machine Learning 2017
Courant Institute of Mathematical Sciences
Homework assignment 1
April 18th, 2017
Due: May 05, 2017

For all these problems, we will adopt the notation used in class.

A. Learning kernels

We consider the scenario of learning kernel.

1. Prove the equivalence of the (primal) optimization problem

$$\min_{\mathbf{w},\mu\in\Delta_q} \frac{1}{2}\sum_{k=1}^{p}\frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C\sum_{i=1}^{m}\max\left\{0, 1-y_i\left(\sum_{k=1}^{p}\mathbf{w}_k\cdot\mathbf{\Phi}_k(x_i)\right)\right\}.$$

and the (dual) problem

$$\max_{\boldsymbol{\alpha}} 2\boldsymbol{\alpha}^\top\mathbf{1} - \left\|\begin{matrix}\boldsymbol{\alpha}^\top\mathbf{Y}^\top\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha}\\ \vdots\\ \boldsymbol{\alpha}^\top\mathbf{Y}^\top\mathbf{K}_p\mathbf{Y}\boldsymbol{\alpha}\end{matrix}\right\|_r$$
$$\text{subject to: } \mathbf{0}\le\boldsymbol{\alpha}\le\mathbf{C}\wedge\boldsymbol{\alpha}^\top\mathbf{y}=0,$$

where $r, q \ge 1$ are conjugate numbers, $\frac{1}{r}+\frac{1}{q}=1$.

Solution: By introducing slack variables $\{\xi_i\}_{i=1}^m$, we can rewrite the primal problem in the following way:

$$\begin{aligned}
&\min_{\mathbf{w},\mu\in\Delta_q} \frac{1}{2}\sum_{k=1}^{p}\frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C\sum_{i=1}^{m}\max\left\{0, 1-y_i\left(\sum_{k=1}^{p}\mathbf{w}_k\cdot\mathbf{\Phi}_k(x_i)\right)\right\}\\
&= \min_{\mu\in\Delta_q}\min_{\mathbf{w}} \frac{1}{2}\sum_{k=1}^{p}\frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C\sum_{i=1}^{m}\max\left\{0, 1-y_i\left(\sum_{k=1}^{p}\mathbf{w}_k\cdot\mathbf{\Phi}_k(x_i)\right)\right\}\\
&= \min_{\mu\in\Delta_q}\min_{\substack{\mathbf{w}\\ \xi_i\ge 1-y_i\left(\sum_{k=1}^{p}\mathbf{w}_k\cdot\mathbf{\Phi}_k(x_i)+b\right),\xi_i\ge 0}} \frac{1}{2}\sum_{k=1}^{p}\frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C\sum_{i=1}^{m}\xi_i
\end{aligned}$$

Consider first the inner minimization problem. Since the inner objective is differentiable and the constraints are affine, the KKT conditions for optimality hold as long as the constraints are feasible (which they are). Thus, we can write the equivalent minimax problem

$$\min_{\mu \in \Delta_q} \min_{\mathbf{w},\xi} \max_{\boldsymbol{\alpha},\boldsymbol{\beta} \geq 0} \frac{1}{2} \sum_{k=1}^{p} \frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C \sum_{i=1}^{m} \xi_i$$
$$+ \sum_{i=1}^{m} \alpha_i \Big(1 - y_i \left(\sum_{k=1}^{p} \mathbf{w}_k \cdot \mathbf{\Phi}_k(x_i) \right) + b - \xi_i \Big) + \sum_{i=1}^{m} \beta_i(-\xi_i),$$

with the associated Lagrangian

$$L(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \sum_{k=1}^{p} \frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C \sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left(\sum_{k=1}^{p} \mathbf{w}_k \cdot \mathbf{\Phi}_k(x_i) \right) + b - \xi_i \right)$$
$$+ \sum_{i=1}^{m} \beta_i(-\xi_i).$$

At optimality, we must have

$$\nabla_{w_k} L = \frac{\mathbf{w}_k}{\mu_k} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{\Phi}_k(x_i) = 0 \Rightarrow \mathbf{w}_k = \mu_k \sum_{i=1}^{m} \alpha_i y_i \mathbf{\Phi}_k(x_i),$$
$$\nabla_b L = \sum_{i=1}^{m} \alpha_i y_i = 0 \Rightarrow \boldsymbol{\alpha}^\top \mathbf{y},$$
$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i + \beta_i = C.$$

Substituting in the equation for C and $\mathbf{w}_k$ shows that at the optimal point, the Lagrangian can be written as

$$\begin{aligned}
L(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\sum_{k=1}^{p} \frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C\sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left(\sum_{k=1}^{p} \mathbf{w}_k \cdot \boldsymbol{\Phi}_k(x_i)\right) + b - \xi_i\right) \\
&\quad + \sum_{i=1}^{m} \beta_i(-\xi_i) \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \mu_k \boldsymbol{\Phi}_k(x_i) \cdot \boldsymbol{\Phi}_k(x_j) \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \mu_k K_k(x_i, x_j) \\
&= \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{y}^\top \left(\sum_{k=1}^{p} \mu_k \mathbf{K}_k\right) \mathbf{y}\boldsymbol{\alpha}.
\end{aligned}$$

By rescaling the expression by 2, we can write the dual problem as:

$$\max_{\substack{0 \le \boldsymbol{\alpha} \le C \\ \boldsymbol{\alpha}^\top \mathbf{y} = 0}} 2\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{y}^\top \left(\sum_{k=1}^{p} \mu_k \mathbf{K}_k\right) \mathbf{y}\boldsymbol{\alpha}.$$

If we now add in the outermost minimization from the original problem, we must solve:

$$\min_{u \in \Delta_q} \max_{\substack{0 \le \boldsymbol{\alpha} \le C \\ \boldsymbol{\alpha}^\top \mathbf{y} = 0}} 2\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{y}^\top \left(\sum_{k=1}^{p} \mu_k \mathbf{K}_k\right) \mathbf{y}\boldsymbol{\alpha}.$$

By Sion's minimax theorem, we can interchange the min and the max and write:

$$\begin{aligned}
&\max_{\substack{0 \le \boldsymbol{\alpha} \le C \\ \boldsymbol{\alpha}^\top \mathbf{y} = 0}} \min_{u \in \Delta_q} 2\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{y}^\top \left(\sum_{k=1}^{p} \mu_k \mathbf{K}_k\right) \mathbf{y}\boldsymbol{\alpha} \\
&= \max_{\substack{0 \le \boldsymbol{\alpha} \le C \\ \boldsymbol{\alpha}^\top \mathbf{y} = 0}} 2\boldsymbol{\alpha}^\top \mathbf{1} - \max_{u \in \Delta_q} \boldsymbol{\alpha}^\top \mathbf{y}^\top \left(\sum_{k=1}^{p} \mu_k \mathbf{K}_k\right) \mathbf{y}\boldsymbol{\alpha}.
\end{aligned}$$

This last expression is linear in μ_k, so we can apply duality of l_p norms to write:

$$\max_{u \in \Delta_q} \boldsymbol{\alpha}^\top \mathbf{y}^\top \mu_k \mathbf{K}_k \mathbf{y}\boldsymbol{\alpha} = \left\| \begin{matrix} \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_1 \mathbf{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K}_p \mathbf{Y}\boldsymbol{\alpha} \end{matrix} \right\|_r,$$

where $\frac{1}{q} + \frac{1}{r} = 1$.

2. What does the problem correspond to for $r = 1$?

   Solution: For $r = 1$ the conjugate variable is $q = \infty$. Thus, the dual problem becomes:

$$\max_{\substack{0 \leq \boldsymbol{\alpha} \leq C \\ \boldsymbol{\alpha}^\top \mathbf{y} = 0}} \min_{u \in \Delta_\infty} 2\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{y}^\top \left(\sum_{k=1}^{p} \mu_k \mathbf{K}_k \right) \mathbf{y}\boldsymbol{\alpha}.$$

   Notice that $\mu_k \geq 0$ implies that $\boldsymbol{\alpha}^\top \mathbf{y}^\top \mu_k \mathbf{K}_k \mathbf{y}\boldsymbol{\alpha} \geq 0$. Moreover, the expression is maximized over $\mu \in \Delta_\infty$ when $\mu_k = 1$ for every $k \in [p]$. Thus, the problem can be written in the following way:

$$\max_{\substack{0 \leq \boldsymbol{\alpha} \leq C \\ \boldsymbol{\alpha}^\top \mathbf{y} = 0}} 2\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{y}^\top \left(\sum_{k=1}^{p} \mathbf{K}_k \right) \mathbf{y}\boldsymbol{\alpha},$$

   which is uniform weight over all the kernels.

B. Deep boosting

Let F be the function defined over $\mathcal{F} = \operatorname{conv}(\bigcup_{k=1}^{p} H_k)$ by

$$F(f) = \widehat{R}_{S,\rho}(f) + \frac{4}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(H_{k_t}),$$

for any $f = \sum_{t=1}^{T} \alpha_t h_t \in \mathcal{F}$. Define the Voted Risk Minimization (VRM) solution as the function f^* minimizing F:

$$f_{\text{VRM}} = \operatorname*{argmin}_{f \in \mathcal{F}} F(f).$$

Let f^* be the element in $\mathcal{F}$ with the smallest generalization error:

$$R(f^*) = \inf_{f \in \mathcal{F}} R(f).$$

1. Fix $\rho > 0$. Use the margin bound presented in class for deep bosting to derive an upper bound on $R(f_{\text{VRM}}) - R_\rho(f^*)$, where $R_\rho(f^*) = \mathrm{E}_{(x,y)\sim D}[1_{yf^*(x) \leq \rho}]$ is the ρ-margin loss of f^*.

Solution: Let $C = \frac{2}{\rho}\sqrt{\frac{\log p}{m}}\left[1 + \sqrt{\left\lceil \log\left[\frac{\rho^2 m}{\log p}\right]\right\rceil}\right]$. Then, the following holds:

$$\begin{aligned}
&\Pr\left[R(f_{\text{VRM}}) - R_\rho(f^*) - \frac{8}{\rho}\sum_{t=1}^T \alpha_t^* \mathfrak{R}_m(H_{k_t}) - 2C > \epsilon\right] \\
&\quad \leq \Pr\left[R(f_{\text{VRM}}) - F(f_{\text{VRM}}) - C > \frac{\epsilon}{2}\right] \\
&\qquad + \Pr\left[F(f_{\text{VRM}}) - R_\rho(f^*) - \frac{8}{\rho}\sum_{t=1}^T \alpha_t^* \mathfrak{R}_m(H_{k_t}) - C > \frac{\epsilon}{2}\right] \\
&\quad \leq 2e^{-\frac{m\epsilon^2}{2}} + \Pr\left[F(f^*) - R_\rho(f^*) - \frac{8}{\rho}\sum_{t=1}^T \alpha_t^* \mathfrak{R}_m(H_{k_t}) - C > \frac{\epsilon}{2}\right] \\
&\quad = 2e^{-\frac{m\epsilon^2}{2}} + \Pr\left[\widehat{R}_S(f^*) - R_\rho(f^*) - \frac{4}{\rho}\sum_{t=1}^T \alpha_t^* \mathfrak{R}_m(H_{k_t}) - C > \frac{\epsilon}{2}\right] \\
&\quad = 2e^{-\frac{m\epsilon^2}{2}} + 2e^{-\frac{m\epsilon^2}{2}} = 4e^{-\frac{m\epsilon^2}{2}}.
\end{aligned}$$

The proof is completed by setting the right-hand side to δ.

2. Compare this result to generalization bound proven in class for SRM. *Solution:* The complexity penalty of the generalization bound for SRM

   was in terms of the hypothesis set of the SRM hypothesis: $\mathfrak{R}_m(H_{k(h^*)})$. In contrast, the complexity term in the VRM bound in the previous question is in terms of a convex combination of the weights of the best ensemble (measured by generalization error). Thus, if the best ensemble does not have too much weight on complex base families, then this term will be much smaller than the SRM term.

   Moreover, the VRM bound has a term that is in $O(\sqrt{\log(p)/m})$ as opposed to the SRM bound which is in $O(\sqrt{\log(k(h^*))/m})$. If we have p hypothesis classes, then the SRM hypothesis will have an index $k(h^*) \leq p$. However, this is not a huge concern since we are taking logarithms of these values.

   Finally, the bound given in the previous question is a margin bound, whereas the SRM bound was a standard generalization bound.

C. Structured prediction

1. Show that $\Phi_u \colon v \mapsto e^{u-v}$ upper bounds $v \mapsto u1_{v\leq 0}$ for all $u \geq 0$.

Solution: When $v > 0$, $u1_{v\leq 0} = 0 \leq e^{u-v}$. When $v \leq 0$, $u1_{v\leq 0} = u \leq e^u \leq e^{u-v}$. □

2. Use that to derive a new structured prediction algorithm based on the hypothesis set

$$\mathcal{H}_2 = \{x \mapsto \mathbf{w} \cdot \mathbf{\Psi}(x, y) \colon \mathbf{w} \in \mathbb{R}^N, \|\mathbf{w}\|_2 \leq \Lambda_2\},$$

for a feature vector $\mathbf{\Psi}$.

Solution: Consider the optimization problem:

$$\min_{\mathbf{w}\in\mathcal{H}_2} \frac{1}{2}\lambda\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \max_{y\neq y_i} \exp\left\{L(y_i, y) - \mathbf{w} \cdot [\mathbf{\Psi}(x_i, y_i) - \mathbf{\Psi}(x_i, y)]\right\}.$$

This objective function is not differentiable. Upper bound max by sum, we have the following optimization problem:

$$\min_{\mathbf{w}\in\mathcal{H}_2} \frac{1}{2}\lambda\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \sum_{y\in\mathcal{Y}} \exp\left\{L(y_i, y) - \mathbf{w} \cdot [\mathbf{\Psi}(x_i, y_i) - \mathbf{\Psi}(x_i, y)]\right\}.$$

3. Assume a bigram feature decomposition $\Psi(x, y) = \sum_{k=1}^{l} \phi(x, k, y_{k-1}, y_k)$. Use that to give an explicit margin bound, assuming that $\|\mathbf{\Psi}\| \leq r$.

Solution: By Theorem 7 in `http://www.cs.nyu.edu/~mohri/pub/vcrf.pdf`,

$$R(h) \leq \widehat{R}^{\text{add}}_{S,\rho}(h) + \frac{4\sqrt{2}}{\rho}\hat{\mathfrak{R}}^{G}_{S}(\mathcal{H}_2) + 3M\sqrt{\frac{\log\frac{1}{\delta}}{2m}},$$

$$R(h) \leq \widehat{R}^{\text{mult}}_{S,\rho}(h) + \frac{4\sqrt{2}M}{\rho}\hat{\mathfrak{R}}^{G}_{S}(\mathcal{H}_2) + 3M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

In addition,

$$\hat{\mathfrak{R}}^{G}_{S}(\mathcal{H}_2) \leq \frac{\Lambda_2 r}{m}\sqrt{\sum_{i=1}^{m}\sum_{f\in F_i}\sum_{y\in\mathcal{Y}_f}|F_i|}.$$

In biagram model, $F_i = \{(y_{i,k-1}, y_{i,k}) : k \in [l]\}$, and $\mathcal{Y}_f = \{(y, y') : y, y' \in \mathcal{Y}\}$. Therefore,

$$\hat{\mathfrak{R}}^{G}_{S}(\mathcal{H}_2) \leq \frac{\Lambda_2 r}{m}\sqrt{ml^2|\mathcal{Y}|^2} = \frac{\Lambda_2 r l|\mathcal{Y}|}{\sqrt{m}}.$$

Combining this with two margin bounds gives the result.

4. Describe in detail an efficient algorithm for the the computation of the gradient for your algorithm.

   Solution: Let

$$F(\mathbf{w}) = \frac{1}{2}\lambda\|\mathbf{w}\|^2 + \sum_{i=1}^{m} F_i(\mathbf{w}),$$

   with

$$F_i(\mathbf{w}) = \sum_{y\in\mathcal{Y}} \exp\left\{L(y_i, y) - \mathbf{w}\cdot[\mathbf{\Psi}(x_i, y_i) - \mathbf{\Psi}(x_i, y)]\right\}.$$

   We need to compute the gradient of $F_i(\mathbf{w})$ efficiently.

$$\begin{aligned}\nabla F_i(\mathbf{w}) = &\sum_{y\in\mathcal{Y}}[\mathbf{\Psi}(x_i, y) - \mathbf{\Psi}(x_i, y_i)]\exp\left\{L(y_i, y) - \mathbf{w}\cdot[\mathbf{\Psi}(x_i, y_i) - \mathbf{\Psi}(x_i, y)]\right\} \\ &= \exp\{-\mathbf{w}\cdot\mathbf{\Psi}(x_i, y_i)\}\underbrace{\left\{\sum_{y\in\mathcal{Y}}\Psi(x_i, y)\exp\{L(y_i, y) + \mathbf{w}\cdot\mathbf{\Psi}(x_i, y)\}\right\}}_{A_i} \\ &- \mathbf{\Psi}(x_i, y_i)\exp\{-\mathbf{w}\cdot\mathbf{\Psi}(x_i, y_i)\}\underbrace{\left\{\sum_{y\in\mathcal{Y}}\exp\{L(y_i, y) + \mathbf{w}\cdot\mathbf{\Psi}(x_i, y)\}\right\}}_{B_i}\end{aligned}$$

   Note that we can easily compute $\exp\{-\mathbf{w}\cdot\mathbf{\Psi}(x_i, y_i)\}$ and $\mathbf{\Psi}(x_i, y_i)\exp\{-\mathbf{w}\cdot\mathbf{\Psi}(x_i, y_i)\}$. Now the problem reduces to computing A_i and B_i efficiently.

   Rewrite A_i and B_i:

$$\begin{aligned}A_i &= \sum_{y\in\mathcal{Y}}\exp\{L(y_i, y) + \mathbf{w}\cdot\mathbf{\Psi}(x_i, y)\}\left(\sum_{k=1}^{l}\phi(x_i, k, y_{k-1}, y_k)\right) \\ &= \sum_{k=1}^{l}\sum_{(s,t)\in\Delta^2}\left[\sum_{\substack{y_{k-1}=s\\ y_k=t}}\exp\{L(y_i, y) + \mathbf{w}\cdot\mathbf{\Psi}(x_i, y)\}\right]\phi(x_k, k, s, t)\end{aligned}$$

$$\begin{aligned} B_i &= \sum_{y \in \mathcal{Y}} \exp\{L(y_i, y) + \mathbf{w} \cdot \mathbf{\Psi}(x_i, y)\} \\ &= \frac{1}{l} \sum_{k=1}^{l} \sum_{(s,t) \in \Delta^2} \left[\sum_{\substack{y_{k-1}=s \\ y_k=t}} \exp\{L(y_i, y) + \mathbf{w} \cdot \mathbf{\Psi}(x_i, y)\} \right] \end{aligned}$$

Therefore the problem further reduces to efficiently computing

$$C_i(k, s, t) = \sum_{\substack{y_{k-1}=s \\ y_k=t}} \exp\{L(y_i, y) + \mathbf{w} \cdot \mathbf{\Psi}(x_i, y)\}$$

for any $k \in [l]$ and $(s, t) \in \Delta^2$.

In what follows, we assume the loss function is decomposable in the same way as bigram feature. That is, for any $y, y' \in \mathcal{Y}$,

$$L(y, y') = \sum_{k=1}^{l} L_k(y_{k-1}, y_k, y'_{k-1}, y'_k).$$

Rewrite $C_i(k, s, t)$ with decomposed loss:

$$\begin{aligned} C_i(k, s, t) = \sum_{\substack{y_{k-1}=s \\ y_k=t}} \Bigg\{ & \prod_{m=1}^{k-1} \exp\{L_m(y_{i,m-1}, y_{i,m}, y_{m-1}, y_m) + \mathbf{w} \cdot \phi(x_i, m, y_{m-1}, y_m)\} \times \\ & \exp\{L_k(y_{i,k-1}, y_{i,k}, s, t) + \mathbf{w} \cdot \phi(x_i, k, s, t)\} \times \\ & \prod_{m=k+1}^{l} \exp\{L_m(y_{i,m-1}, y_{i,m}, y_{m-1}, y_m) + \mathbf{w} \cdot \phi(x_i, m, y_{m-1}, y_m)\} \Bigg\}. \end{aligned}$$

Now we see $C_i(k, s, t)$ is the sum of the weights of all paths going through a given transition (s, t). We can use the flow computation introduced in class to efficiently compute it. Once $C_i(k, s, t)$s are computed, we can easily get A_i, B_i, and finally $\nabla F(\mathbf{w})$.

Bonus: Multi-Armed Bandit

Consider the standard multi-armed bandit problem with N arms, but assume now that the learner receives extra information as follows. Assume that there is a function $O : \{1, 2, \ldots, N\} \to 2^{\{1,2,\ldots,N\}}$ (where 2^A indicates the power set of A) such that at any round, pulling arm i results in knowledge of the rewards of all arms in $O(i)$. Assume further that $i \in O(i)$ (i.e. pulling an arm always results in knowledge of that arm's reward) and $i \in O(j) \Leftrightarrow j \in O(i)$.

1. Explain how the function O induces an undirected graph $G = (V, E)$ over the arms of the game.

2. Recall that a *clique* C of a graph is a subset of its vertices such that every vertex is connected to every other vertex in this subset. Moreover, a *clique covering* of a graph is a set of cliques such that their union is equal to the entire set of vertices in the graph. Let $\mathcal{C}$ be a clique covering of the graph described in the previous question, so that $\cup_{C \in \mathcal{C}} C = V$. Design an algorithm that achieves the following regret bound:

$$T\mu_* - \sum_{t=1}^{T} \mu_{I_t} \leq \mathcal{O}\left(\inf_{\mathcal{C}} \left\{ \sum_{C \in \mathcal{C}} \frac{\max_{i \in C} \Delta_i \log(T)}{\min_{j \in C} \Delta_j^2} \right\}\right).$$

3. Explain how this regret bound compares to results for the standard MAB problem and for the full-information setting.

Solution: See `http://www.auai.org/uai2012/papers/236.pdf`.