Kernel Learning with Non-convex Ramp Loss

Xingke Chen
College of Science
China University of Petroleum
chenxk1229@hotmail.com

Xijun Liang
College of Science
China University of Petroleum

Ling Jian
School of Economics and Management
China University of Petroleum
bebetter@upc.edu.cn

Abstract

We study the kernel learning problems with ramp loss, a non-convex but noise-resistant loss function. In this work, we justify the validity of ramp loss under classical kernel learning framework, in particular, we show that the generalization bound for empirical ramp risk minimizer is similar to that of convex surrogate losses, which implies kernel learning with such loss function is not only noise resistant but, more important, statistically consistent. For adapting to large-scale and real-time scenarios, we introduce PA-RAMP, a heuristic online algorithm based on passive aggressive framework, to solve this learning problem. Empirically, with less support vectors, this algorithm achieves comparable empirical performances to varied online methods.

1 Introduction

Kernel learning methods based on convex loss functions, including hinge loss, squared hinge loss and cross entropy etc., have made a great success in the past decades. An important reason is that for such problems there exist abundant theoretical results and efficient optimization methods [27, 10, 6, 4]. However, machine learning problems with corruptions and noisy or even adversarial data are extremely common in practice [11], such as datasets with wrong labels, and PU learning, where unlabeled samples are compulsorily labeled as negative samples then trained by binary classification algorithms [23], etc. In these cases the convex loss-based models usually have bad performances [21], albeit they have desirable statistical and optimization properties.

Robust losses are believed to be feasible for these problems [13] since they are able to make the models insensitive to the data by weakening the influence of the extreme data points. As a robust loss function, ramp loss can also induce sparse solutions [7] due to the non-convexity, leading to lightweight models that are robust against the noise. Therefore, it is widely used in many applications such as anomaly detection, machine translation, and scene recognition [26, 28, 15, 18] etc. Despite such a success, there are some facts that discourage the use of ramp loss:

- First, unlike convex loss functions, the performance of empirical ramp risk minimizer can not be guaranteed, more precisely, it is unclear whether minimizing the empirical risk under ramp loss can also minimize the misclassification rate.
- Second, in algorithmic perspective, the Concave Convex Procedure proposed by [7] tries to approximate the empirical ramp risk minimizer by solving quadratic programming iteratively. In many large-scale and real-time learning scenarios, such as click-through rate,

Preprint. Under review.

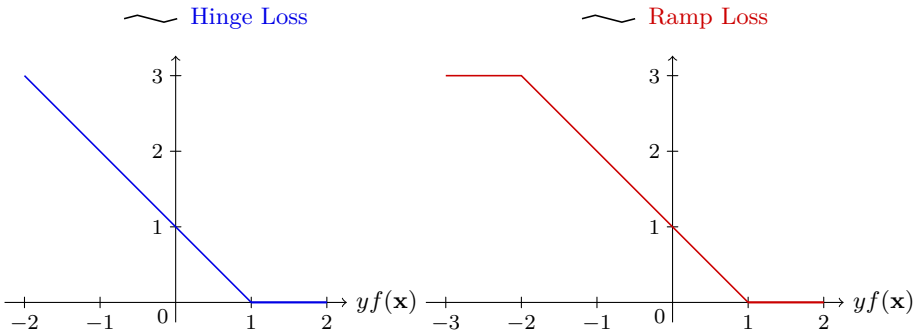


Figure 1: The comparison between hinge loss (left) and ramp loss with $s = -2$ (right). Hinge loss is a special case of ramp loss with $s \to -\infty$.

recommendation system and so on [22, 1, 5], this kind of batch method is not scalable and can be prohibitively slow.

In this paper, we develop the learning theory for ramp loss and introduce PA-RAMP, an online algorithm for approximate the empirical risk minimizer in large-scale or data streaming applications. Specifically, we start by presenting the basic notions of this problem in Section 2. Then in Section 3, following the methodology in [4], we show that ramp loss satisfies the classification-calibrated condition, which implies the risk minimizer under ramp loss is faithful to the Bayes classifier[1]. In addition, we analyze the statistical properties of ramp loss under kernel learning framework, as a result, the excess risk of empirical ramp risk minimizer has the rate $O\left(\frac{1}{\sqrt{n}}\right)$, implying kernel learning with ramp loss is theoretically guaranteed. In Section 4 and Section 5, we discuss the online algorithm PA-RAMP and its empirical performance on multiple moderate and large-scale datasets. Finally we end up with the conclusion.

2 Preliminaries

Given i.i.d. training data $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ from a distribution $\mathbb{P}$ over $\mathcal{X} \times \{-1, +1\}$, based on these empirical data, a classification algorithm tries to find the Bayes classifier, the classifier achieves minimal misclassification rate,

$$f^\star \in \arg\min_{f:\mathcal{X}\to\mathbb{R}} R(f) := \arg\min_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{(X,Y)\sim\mathbb{P}} \left[\mathbb{I}\left[Yf(X) < 0\right]\right] = \arg\min_{f:\mathcal{X}\to\mathbb{R}} \mathbb{P}[Yf(X) < 0]. \tag{1}$$

It is well-known that Bayes classifier has the form

$$\operatorname{sign}(f^\star) : x \mapsto 2\mathbb{I}\left[\eta(x) \geq \frac{1}{2}\right] - 1, \tag{2}$$

where $\mathbb{I}\left[\cdot\right]$ is the indicator function and $\eta(x) = \mathbb{P}\left[Y = +1 \mid X = x\right]$.

However, the optimization problem above is computationally intractable, we usually replace the original $0-1$ loss function with some surrogate function $\phi : \mathbb{R} \to \mathbb{R}_+$. Surrogate functions are convex in general for computational consideration, for example, cross entropy, hinge loss and exponential loss [14, 8, 12] are common choices. In this paper, the surrogate is ramp loss

$$\phi_s\left(yf\left(x\right)\right) = \min\left\{\max\left\{1 - yf\left(x\right), 0\right\}, 1 - s\right\}, \quad (s \leq 0), \tag{3}$$

where s is a hyperparameter that controls the extent of noise resistance (see Figure 1). With this notation, for a specific classifier f, the risk and empirical risk with respect to ramp loss are defined by

$$R_{\phi_s}(f) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[\phi_s\left(Yf(X)\right)\right], \quad \hat{R}_{\phi_s}(f) = \frac{1}{n}\sum_{i=1}^{n} \phi_s\left(Y_i f(X_i)\right) \tag{4}$$

[1]After finishing this work, we learned that Huang et al. [17] obtained a similar result as well.

accordingly. Under such risk measure, the minimum achievable risk under ramp loss can be written as

$$R_{\phi_s}^{\star} = \min_{f:\mathcal{X}\to\mathbb{R}} R_{\phi_s}(f). \tag{5}$$

If we restrict the classifier f in a reproducing kernel Hilbert space (RKHS), then the kernel learning method under ramp loss can be formularized as

$$\hat{f}_n^{\phi_s} = \arg\min_{f\in\mathcal{H}} \hat{R}_{\phi_s}(f), \quad \mathcal{H} = \left\{ f \in \mathcal{H}_{\mathcal{K}} \mid \|f\|_{\mathcal{H}_{\mathcal{K}}} \leq \lambda \right\} \tag{6}$$

for some regularization coefficient $\lambda > 0$, where $\mathcal{H}_{\mathcal{K}}$ is an RKHS defined by a positive semidefinite kernel $\mathcal{K}$ with norm $\|\cdot\|_{\mathcal{H}_{\mathcal{K}}}$ (see [2] for details) and we call $\hat{f}_n^{\phi_s}$ the empirical risk minimizer under ramp loss. The statistical properties of $\hat{f}_n^{\phi_s}$ are appealing since it has been widely used previously [7, 15, 28], in the next section, we will show its risk consistency, namely,

$$R(\hat{f}_n^{\phi_s}) \xrightarrow{\mathbb{P}} R(f^{\star}). \tag{7}$$

3 Main theoretical results

In this section, we develop the learning theory for kernel learning with ramp loss. The main results are two-fold: first, by showing that the ϕ_s-risk minimizer (i.e. $\arg\min_{f:\mathcal{X}\to\mathbb{R}} R_{\phi_s}(f)$) agrees with Bayes classifier $f^{\star}$ pointwisely, we justify the validity of using ramp loss. Second, we obtain the generalization bound for $\hat{R}_s$ under kernel learning framework, which shows the learning problem Equation 6 is guaranteed.

Under ramp loss ϕ_s, f achieves minimum ϕ_s-risk $R_{\phi_s}^{\star}$ if it minimizes the following conditional risk

$$C_{\eta(x)}(f(x)) := \mathbb{E}\left[\phi_s\left(Yf(X)\right) \mid X = x\right] = \eta\left(x\right)\phi_s\left(f(x)\right) + \left(1-\eta\left(x\right)\right)\phi_s\left(-f(x)\right) \tag{8}$$

at each x. For classification, we say a loss is reasonable if the risk minimizer under the loss function is Bayes classifier $f^{\star}$. This kind of sense is captured by the following classification calibration condition coined by [4]. For convenience, we will replace $(\eta(x), f(x))$ with (η, α) when we talk about the conditional probability and prediction at a specific point x.

Definition 1 (Classification calibrated loss)**.** *A loss function ϕ is classification calibrated if for any $\eta \neq \frac{1}{2}$,*

$$H^{-}\left(\eta\right) > H\left(\eta\right),$$

where $H\left(\eta\right) = \inf_{\alpha\in\mathbb{R}} C_{\eta}(\alpha)$ and $H^{-}\left(\eta\right) = \inf_{\alpha:\alpha(2\eta-1)\leq 0} C_{\eta}(\alpha)$.

This condition is fairly concrete: for classification-calibrated loss ϕ, the prediction α of data point x that does not agree with Bayes classifier at that point will incur larger conditional ϕ-risk, therefore, the ϕ-risk minimizer will alwalys agree with Bayes classifier at all x. The following Theorem 1 shows that ramp loss satisfies the classification-calibrated condition.

Theorem 1. *Ramp loss is classification-calibrated.*

As we discussed earlier, such a result guarantees the minimizer w.r.t. ramp loss is faithful to Bayes classifier and hence it is reasonable to use ramp loss in practice. However, it does not give us quantitative information, in other words, we need finite-sample results to show learning with ramp loss is theoretically sound. And we show this in the following sequel.

Theorem 2 bridges the gap between excess risks in terms of misclassification rate and ϕ_s-risk. It reveals the property of ramp loss: if $\hat{f}_n^{\phi_s}$ achieves minimum ϕ_s-risk, then it also achieves Bayes risk.

Theorem 2. *For ramp loss ϕ_s, if $R_{\phi_s}(\hat{f}_n^{\phi_s}) \xrightarrow{\mathbb{P}} R_{\phi_s}^{\star}$, the optimal expected risk with respect to ramp loss, then $R(\hat{f}_n^{\phi_s}) \xrightarrow{\mathbb{P}} R(f^{\star})$. In particular, we have*

$$R(\hat{f}_n^{\phi_s}) - R(f^{\star}) \leq R_{\phi_s}(\hat{f}_n^{\phi_s}) - R_{\phi_s}^{\star}.$$

With Theorem 2, we can prove the generalization bound of $\hat{f}_n^{\phi_s}$ under kernel learning framework in terms of misclassification rate.

Theorem 3. *Let $\mathcal{H}_\mathcal{K}$ be a reproducing kernel Hilbert space with kernel $\mathcal{K}$ such that $\sup_{x\in\mathcal{X}} |\mathcal{K}(x,x)| \leq k$. Under ramp loss ϕ_s, the empirical ϕ_s-risk minimizer $\hat{f}_n^{\phi_s}$ over $\mathcal{H} = \left\{f \in \mathcal{H}_\mathcal{K} \mid \|f\|_{\mathcal{H}_\mathcal{K}} \leq \lambda^\star\right\}$ satisfies the following generalization bound:*

$$R(\hat{f}_n^{\phi_s}) - R(f^\star) \leq \inf_{f\in\mathcal{H}} R_{\phi_s}(f) - R^\star_{\phi_s} + \left(8 + 2\sqrt{2\log\frac{2}{\delta}}\right)\lambda^\star\sqrt{\frac{k}{n}}$$

with probability $1-\delta$. Moreover, if the ramp risk minimizer is in $\mathcal{H}$, that is,

$$\inf_{f\in\mathcal{H}} R_{\phi_s}(f) = R^\star_{\phi_s},$$

then $\hat{f}_n^{\phi_s}$ is risk consistent (Equation (7)).

Theorem 3 gives us the quantitative theoretical guarantee of kernel learning with ramp loss. Moreover, it reveals that the estimator under ramp loss has the similar convergence rate as the estimators under convex surrogates.

Algorithm 1 PA-RAMP

Input: Training set $\mathcal{S} = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, kernel $\mathcal{K}$, loss function parameter $s \leq 0$.
Output: Classifier f_n.

```
Initialize the empty list A for coefficients, and empty list K for support vectors, f_0 = 0.
for i = 1 to n do
    Receive the data (X_i, Y_i).
    Compute the ramp loss ℓ_i = φ_s(Y_i f(X_i)).
    if 0 < ℓ_i < 1 − s then
        Append X_i to K.
        Append α_i = ℓ_i Y_i / K(X_i, X_i) to A.
        f_i ← Σ_{l=1}^{|A|} A[l] K(K[l], ·), where |A| is the current length of the list A.
    end if
end for
```

4 Optimization

We have discussed the statistical consistency of the kernel learning method with ramp loss, such a property holds if $\hat{f}_n^{\phi_s}$ can be obtained exactly. However, due to the non-convexity of ramp loss, existing methods [7] are only guaranteed to converge to a local minimum and are not scalable as the sample size increasing. Instead of finding the global minimum, in this work, we formulate an online algorithm PA-RAMP that is motivated by the online passive aggressive framework [9] and appropriate for large-scale or data streaming learning problems.

PA-RAMP approximates

$$\hat{f}_n^{\phi_s} = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \phi_s\left(Y_i f(X_i)\right) \tag{9}$$

with a sequence of classifiers $\{f_0, f_1, \ldots, f_n\}$ in the space $\mathcal{H} = \left\{f \in \mathcal{H}_\mathcal{K} \mid \|f\|_{\mathcal{H}_\mathcal{K}} \leq \lambda^\star\right\}$. This algorithm processes the instances one by one, on each round (say i), it observes an instance X_i and predicts the label, then it suffers ramp loss when the true label Y_i is revealed. As a result, the algorithm updates the model on the basis of the ramp loss at (X_i, Y_i).

The update scheme of PA-RAMP is:

- Begin with the initial point $f_0 = 0$.
- At i-th round, if $0 < \phi_s(Y_i f_{i-1}(X_i)) < 1 - s$, the algorithm aggressively forces f_i to achieve zero loss at X_i and lets f_i be close to f_{i-1} as much as possible. That is,

$$f_i = \arg\min_{f\in\mathcal{H}} \frac{1}{2}\|f - f_{i-1}\|^2_{\mathcal{H}_\mathcal{K}}, \quad \text{s.t. } Y_i f(X_i) \geq 1.$$

Table 1: The dataset descriptions used in the experiments

Name	Number of instances	Number of features
Breast Cancer	569	32
Letters	20000	16
Occupancy	20560	7
Banana	5300	2
Adult	48842	14

The Lagrangian is given by

$$\mathcal{L}(f,\tau) = \frac{1}{2}\left\|f - f_{i-1}\right\|_{\mathcal{H}_\mathcal{K}}^2 - \tau(Y_i f(X_i) - 1), \quad \tau \geq 0.$$

According to Karush-Kuhn-Tucker conditions

$$\begin{cases} \frac{\partial \mathcal{L}(f,\tau)}{\partial f}\Big|_{f=f_i} = f_i - f_{i-1} - \tau Y_i \frac{\partial f(X_i)}{\partial f} = f_i - f_{i-1} - \tau Y_i \mathcal{K}(X_i, \cdot) = 0 \\ \frac{\partial \mathcal{L}(f_i,\tau)}{\partial \tau} = -Y_i f_i(X_i) + 1 = -Y_i \left(f_{i-1}(X_i) + \tau Y_i \mathcal{K}(X_i, X_i)\right) + 1 = 0, \end{cases}$$

we have

$$f_i = f_{i-1} + \frac{Y_i \phi_s(Y_i f_{i-1}(X_i))}{\mathcal{K}(X_i, X_i)} \mathcal{K}(X_i, \cdot).$$

- Otherwise, the algorithm is passive and does not update.

PA-RAMP is summarized in Algorithm 1.

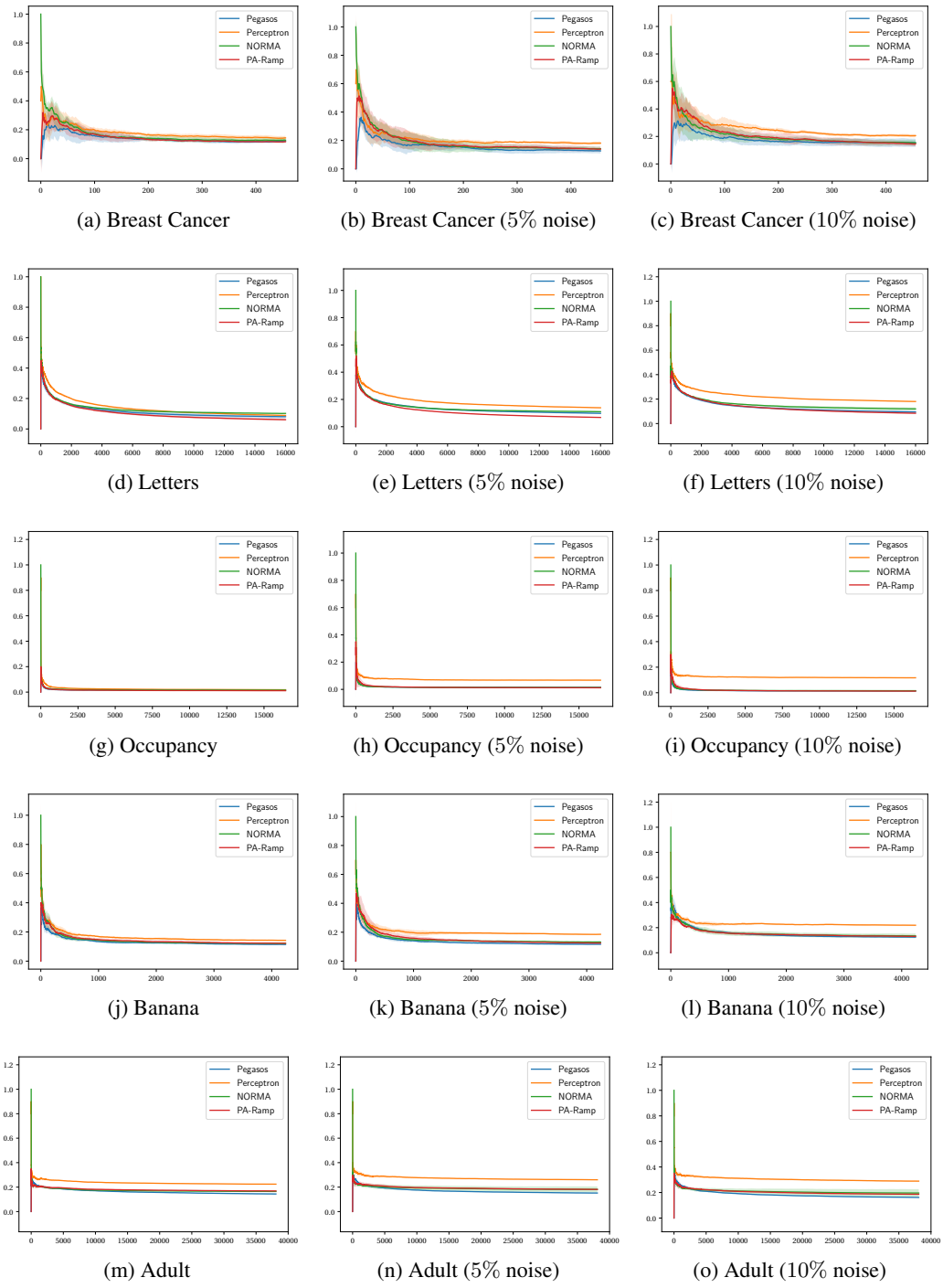
5 Experiments

In this section, we compare the performances of PA-RAMP with other online learning algorithms that using kernel method: Perceptron, Pegasos, as well as NORMA [25, 19, 24]. The datasets are collected from the UCI Machine Learning Repository [3] and the descriptions of the datasets are given in Table 1. All datasets are binary classification datasets but Letters dataset, which is casted to be a binary one by grouping the first 13 letters and the other part to two classes. Moreover, we drop the instances with missing value and standardize all datasets before the experiments.

All experiments in this section are conducted in Python 2.7.13 environment on Windows 10 server 2012 R2 system with Intel(R) Xeon(R) E5-2640 processors and 64G RAM. The kernel function is Gaussian kernel with parameter 1 and the hyperparameters of all algorithms are tuned by 5-fold cross-validation via a hold-out set with 1000 instances (for Breast Cancer dataset, 10% instances), where we choose the ramp loss parameter s in $[-1, -0.75, -0.5, \ldots, 0]$, the learning rates η for corresponding algorithms in $[10^{-5}, 10^{-4}, \ldots, 10^{-1}]$, truncated parameter τ of NORMA in $[10^{-5}, 10^{-4}, 10^{-3}]$, and regularization coefficients (for Pegasos and NORMA) λ in $[10^{-3}, 10^{-2}, \ldots, 1]$. With those tuned hyperparameters, we report the performances of all algorithms by averaging on 5 random train-test splits with 80% training data and 20% testing data. To explore the performances of those methods under noisy scenarios, we flip the labels of training data randomly with probabilities 0%, 5% and 10%.

The results in Table 2 include the accuracies on test sets and the number of support vectors (along with standard deviations), and Figure 2 demonstrates how misclassification rates decrease with the increase of iterations. Getting benefits from ramp loss, the proposed PA-RAMP induces significantly less support vectors than others except perceptron, which only identifies misclassified observations as support vectors and hence has unstable and unsatisfied performances. Moreover, with fewer support vectors, PA-RAMP achieves comparable and even better performances to others on multiple datasets. Under noisy scenarios, there is no enough evidence shows that PA-RAMP is superior to others overwhelmingly, but we observe that as the increase of noisy level, other methods tend to store more support vectors, while PA-RAMP is more stable against noise.

To conclude, PA-RAMP is able to store less parameters and spends less time while predicting, which is crucial in large-scale learning or real-time prediction problems. In noise case, PA-RAMP has similar performances to other methods with more stable number of support vectors. Therefore, in large-scale learning scenarios, the proposed algorithm is preferred among these methods.



(a) Breast Cancer (b) Breast Cancer (5% noise) (c) Breast Cancer (10% noise)

(d) Letters (e) Letters (5% noise) (f) Letters (10% noise)

(g) Occupancy (h) Occupancy (5% noise) (i) Occupancy (10% noise)

(j) Banana (k) Banana (5% noise) (l) Banana (10% noise)

(m) Adult (n) Adult (5% noise) (o) Adult (10% noise)

Figure 2: Numbers of iterations versus misclassification rates of different datasets.

Table 2: Numerical experiments: PA-RAMP versus other algorithms

Breast Cancer	Online Methods							
	Perceptron		Pegasos		NORMA		PA-RAMP	
	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs
Noise level 0%	88.42 ± 7.07	65.6 ± 6.1	89.12 ± 2.75	105.2 ± 17.8	92.11 ± 2.15	138.4 ± 24.1	92.28 ± 2.03	113.0 ± 11.1
Noise level 5%	87.37 ± 4.49	99.8 ± 6.8	88.42 ± 6.11	121.8 ± 19.1	88.60 ± 3.86	184.0 ± 34.4	89.65 ± 1.40	115.8 ± 11.7
Noise level 10%	84.91 ± 4.35	119.2 ± 8.4	87.37 ± 5.13	176.2 ± 41.7	86.84 ± 3.42	205.6 ± 34.5	87.71 ± 3.19	106.0 ± 8.4
Letters	Online Methods							
	Perceptron		Pegasos		NORMA		PA-RAMP	
	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs
Noise level 0%	94.70 ± 0.23	1419.6 ± 10.3	93.62 ± 1.11	9344.8 ± 914.9	91.45 ± 0.32	14560.0 ± 2868.0	96.72 ± 0.23	7473.0 ± 210.9
Noise level 5%	89.27 ± 0.79	2811.0 ± 44.3	90.46 ± 0.68	10130.4 ± 56.0	90.47 ± 1.11	13255.4 ± 3362.7	96.52 ± 0.21	7852.8 ± 287.6
Noise level 10%	84.63 ± 1.17	3919.8 ± 45.6	91.73 ± 0.86	9379.0 ± 594.9	89.72 ± 2.17	14181.8 ± 3638.4	95.44 ± 0.73	7574.8 ± 616.0
Occupancy	Online Methods							
	Perceptron		Pegasos		NORMA		PA-RAMP	
	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs
Noise level 0%	98.78 ± 0.20	303.6 ± 5.9	98.74 ± 0.17	5837.6 ± 2798.1	98.64 ± 0.19	470.0 ± 230.6	99.00 ± 0.12	309.2 ± 15.5
Noise level 5%	96.51 ± 0.63	1833.4 ± 83.2	98.64 ± 0.68	7593.2 ± 2491.2	98.73 ± 0.16	8353.2 ± 4770.2	98.97 ± 0.09	306.6 ± 10.7
Noise level 10%	87.55 ± 4.66	3221.8 ± 42.8	98.56 ± 0.16	8175.4 ± 2184.1	98.65 ± 0.13	8684.4 ± 4536.0	98.92 ± 0.03	340.4 ± 36.7
Banana	Online Methods							
	Perceptron		Pegasos		NORMA		PA-RAMP	
	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs
Noise level 0%	86.17 ± 1.80	602.4 ± 5.6	89.98 ± 0.70	1554.4 ± 8.9	89.38 ± 1.80	1434.0 ± 230.7	89.77 ± 0.95	643.0 ± 27.5
Noise level 5%	85.23 ± 9.46	927.2 ± 18.3	88.94 ± 0.41	1696.2 ± 39.4	83.91 ± 1.80	1873.0 ± 439.1	88.49 ± 0.89	685.0 ± 26.2
Noise level 10%	82.04 ± 5.66	1173.2 ± 20.1	88.92 ± 0.62	1894.4 ± 25.0	87.66 ± 3.09	2130.2 ± 230.6	87.34 ± 2.83	692.0 ± 47.2
Adult	Online Methods							
	Perceptron		Pegasos		NORMA		PA-RAMP	
	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs	Acc(%)	#SVs
Noise level 0%	78.83 ± 0.36	8537.2 ± 24.9	84.15 ± 0.23	23632.2 ± 543.6	84.26 ± 0.27	27780.2 ± 598.2	84.09 ± 0.36	19287.0 ± 1476.9
Noise level 5%	75.45 ± 0.43	10808.8 ± 80.9	83.74 ± 0.55	24092.0 ± 37.2	82.09 ± 3.56	30657.2 ± 14870.6	83.32 ± 0.56	21165.8 ± 1805.1
Noise level 10%	72.10 ± 0.29	12664.4 ± 118.5	83.30 ± 0.25	23996.2 ± 260.0	81.11 ± 3.50	23962.2 ± 17350.7	83.09 ± 0.33	19588.2 ± 1374.8

6 Conclusion

In this paper, we study the non-convex ramp loss that is widely used in noisy scenarios. We justify the validity of using ramp loss and provide the risk generalization bound for the empirical ramp risk minimizer under kernel learning framework. For large-scale and real-time problems, we introduce PA-RAMP algorithm. With much fewer support vectors, its empirical performances in such problems are competitive with other online algorithms based on kernel method.

Appendix

Proofs of Theorem 1

Consider any $x \in \mathcal{X}$ and denote $(\eta(x), f(x))$ by (η, α). Then the conditional risk under ramp loss has the following form:

$$\begin{aligned} C_\eta(\alpha) &= \eta\phi_s(\alpha) + (1-\eta)\phi_s(-\alpha) \\ &= \min\{\max\{0, \eta(1-\alpha)\}, \eta(1-s)\} + \min\{\max\{0, (1-\eta)(1+\alpha)\}, (1-\eta)(1-s)\}. \end{aligned}$$

Note that $C_\eta(\alpha)$ has different forms under different s.

- When the parameter s satisfies $s < -1$,

$$C_\eta(\alpha) = \begin{cases} \eta(1-s) & \alpha \le s \\ \eta(1-\alpha) & s < \alpha \le -1 \\ (1-2\eta)\alpha + 1 & -1 < \alpha \le 1 \\ (1-\eta)(1+\alpha) & 1 < \alpha \le -s \\ (1-\eta)(1-s) & \alpha > -s. \end{cases}$$

  Thus $H(\eta) = 2(1-\eta) < 1 = H^-(\eta)$ for $\eta > \frac{1}{2}$ and $H(\eta) = 2\eta < 1 = H^-(\eta)$ for $\eta < \frac{1}{2}$, which implies ramp loss is classification calibrated when $s < -1$.

- For $-1 \le s \le 0$,

$$C_\eta(\alpha) = \begin{cases} \eta(1-s) & \alpha \le -1 \\ 1 + (1-\eta)\alpha - \eta s & -1 < \alpha \le s \\ 1 + \alpha(1-2\eta) & s < \alpha \le -s \\ 1 + (\eta-1)s - \eta\alpha & -s < \alpha \le 1 \\ (1-\eta)(1-s) & \alpha > 1. \end{cases}$$

  Thus $H(\eta) = (1-\eta)(1-s) < \min\{\eta(1-s), 1\} = H^-(\eta)$ for $\eta > \frac{1}{2}$ and $H(\eta) = \eta(1-s) < (1-\eta)(1-s) = H^-(\eta)$ for $\eta < \frac{1}{2}$, which implies ramp loss is classification calibrated when $-1 \le s \le 0$.

Thus, we have proved that ramp loss is classification calibrated.

Proofs of Theorem 2

To proof Theorem 2, we need the following fundamental fact that was proved in [4].

Lemma 1. *For any non-negative loss function ϕ, any measurable function $f : \mathcal{X} \to \mathbb{R}$ and any distribution $\mathbb{P}$ over $\mathcal{X} \times \{-1, +1\}$, there is a function $\psi : [-1, 1] \to [0, \infty)$, which is the biconjugate function of*

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

such that

$$\psi\left(R(f) - R(f^\star)\right) \le R_\phi(f) - R_\phi^\star.$$

Since ramp loss ϕ_s satisfies the conditions of Lemma 1. By virtue of the Proofs of Theorem 1, we can derive the closed form of $\tilde{\psi}$ for different s (see Figure 3).

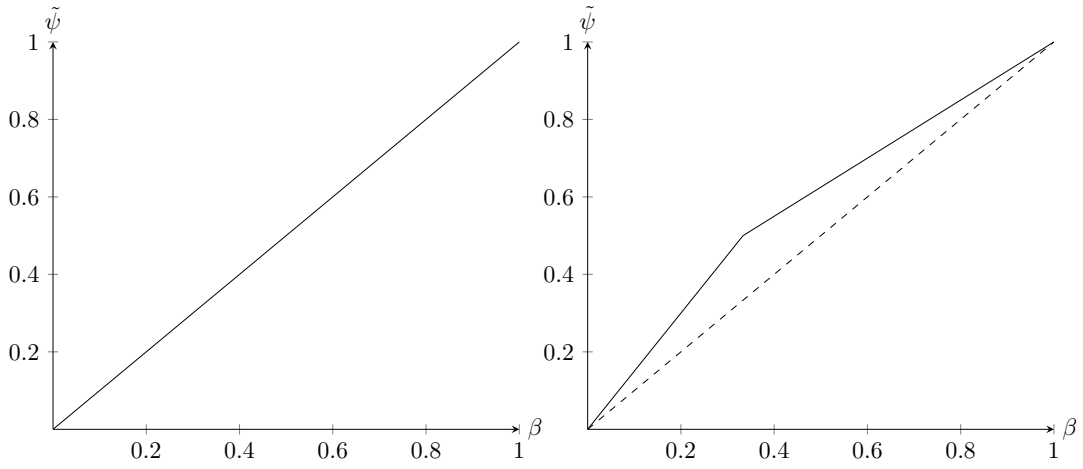


Figure 3: The graph of $\tilde{\psi}$ on $[0,1]$ for $s<-1$ (left) and $-1\le s\le 0$ (right), and the dotted line is the biconjugation on $[0,1]$.

- Suppose $s<-1$,

$$\tilde{\psi}(\beta)=H^-\left(\frac{1+\beta}{2}\right)-H\left(\frac{1+\beta}{2}\right)=1-\min\{1+\beta,1-\beta\}=\begin{cases}\beta, & \beta>0\\ 0, & \beta=0\\ -\beta, & \beta<0.\end{cases}=|\beta|.$$

  Due to the convexity and closedness of $\tilde{\psi}$, we have $\psi=\tilde{\psi}$, which is a continuous and monotone increasing function on $[0,1]$, and hence its inverse function exists and is continuous. By Lemma 1 and continuous mapping theorem, if $R_{\phi_s}(\hat{f}_n^{\phi_s})\xrightarrow{\mathbb{P}}R_{\phi_s}^\star$, then $R(\hat{f}_n^{\phi_s})\xrightarrow{\mathbb{P}}R(f^\star)$. In addition, we can obtain the convergence rate if the convergence rate of the RHS is known because

$$R(\hat{f}_n^{\phi_s})-R(f^\star)\le\psi^{-1}\left(R_{\phi_s}(\hat{f}_n^{\phi_s})-R_{\phi_s}^\star\right)=R_{\phi_s}(\hat{f}_n^{\phi_s})-R_{\phi_s}^\star,$$

  where the equality holds due to $R_{\phi_s}(\hat{f}_n^{\phi_s})-R_{\phi_s}^\star\ge 0$.

- Suppose $-1\le s\le 0$, plug $\frac{1+\beta}{2}$ into the expression of H and H^- in the Proofs of Theorem 1,

$$\begin{aligned}\tilde{\psi}(\beta)&=H^-\left(\frac{1+\beta}{2}\right)-H\left(\frac{1+\beta}{2}\right)\\&=\begin{cases}1-\frac{1-\beta}{2}(1-s), & \beta>\frac{2}{1-s}-1\\ \beta(1-s), & 0<\beta\le\frac{2}{1-s}-1\\ 0, & \beta=0\\ -\beta(1-s), & \beta<0.\end{cases}.\end{aligned}$$

  Since the epigraph of the biconjugate function ψ is the closed convex hull of the epigraph of $\tilde{\psi}$ (see [16]), we have

$$\psi(\beta)=\begin{cases}-(1-s)\beta, & \beta\in[-1,0)\\ \beta, & \beta\in[0,1]\end{cases}.$$

  Since $R_{\phi_s}(\hat{f}_n^{\phi_s})-R_{\phi_s}^\star\ge 0$, we have

$$R(\hat{f}_n^{\phi_s})-R(f^\star)\le\psi^{-1}\left(R_{\phi_s}(\hat{f}_n^{\phi_s})-R_{\phi_s}^\star\right)=R_{\phi_s}(\hat{f}_n^{\phi_s})-R_{\phi_s}^\star.$$

Proofs of Theorem 3

By the result in Theorem 2, we have

$$\begin{aligned}
R(\hat{f}_n^{\phi_s}) - R(f^\star) &\leq R_{\phi_s}(\hat{f}_n^{\phi_s}) - R_{\phi_s}^\star \\
&= R_{\phi_s}(\hat{f}_n^{\phi_s}) - \hat{R}_{\phi_s}(\hat{f}_n^{\phi_s}) + \hat{R}_{\phi_s}(\hat{f}_n^{\phi_s}) - \hat{R}_{\phi_s}(\tilde{f}) + \hat{R}_{\phi_s}(\tilde{f}) - \inf_{f\in\mathcal{H}} R_{\phi_s}(f) + \inf_{f\in\mathcal{H}} R_{\phi_s}(f) - R_{\phi_s}^\star \\
&\leq \inf_{f\in\mathcal{H}} R_{\phi_s}(f) - R_{\phi_s}^\star + 2\sup_{f\in\mathcal{H}} \left|\hat{R}_{\phi_s}(f) - R_{\phi_s}(f)\right|,
\end{aligned}$$

where $\tilde{f} = \arg\min_{f\in\mathcal{H}} R_{\phi_s}(f)$. Define $Z_i = (X_i, Y_i)$ and

$$g(Z_1, Z_2, \ldots, Z_n) = \sup_{f\in\mathcal{H}} \left|\hat{R}_{\phi_s}(f) - R_{\phi_s}(f)\right|,$$

for any $z_1, z_2, \ldots, z_n$ and z_i',

$$|g(z_1, \ldots z_i, \ldots, z_n) - g(z_1, \ldots z_i', \ldots, z_n)| \leq \frac{1}{n}\sup_{f\in\mathcal{H}} |\phi_s(y_i' f(x_i')) - \phi_s(y_i f(x_i))|.$$

Since ϕ_s is monotone and for any $f \in \mathcal{H}$, according to reproducing property of f,

$$\sup_{x\in\mathcal{X}} |f(x)| = \sup_{x\in\mathcal{X}} \left|\langle \mathcal{K}(x,\cdot), f\rangle_{\mathcal{H}_\mathcal{K}}\right| \leq \|f\|_{\mathcal{H}_\mathcal{K}} \sup_{x\in\mathcal{X}} \|\mathcal{K}(x,\cdot)\|_{\mathcal{H}_\mathcal{K}} \leq \lambda^\star\sqrt{k},$$

we have the following bounded difference condition

$$|g(z_1, \ldots z_i, \ldots, z_n) - g(z_1, \ldots z_i', \ldots, z_n)| \leq \frac{1}{n}\left|\phi_s\left(\lambda^\star\sqrt{k}\right) - \phi_s(-\lambda^\star\sqrt{k})\right| \leq \frac{2\lambda^\star\sqrt{k}}{n},$$

where the last inequality holds because ϕ_s is 1-Lipschitz. Therefore g can be bounded via bounded difference inequality:

$$\mathbb{P}\left[\left|\sup_{f\in\mathcal{H}} \left|\hat{R}_{\phi_s}(f) - R_{\phi_s}(f)\right| - \mathbb{E}\left[\sup_{f\in\mathcal{H}} \left|\hat{R}_{\phi_s}(f) - R_{\phi_s}(f)\right|\right]\right| \geq t\right] \leq 2\exp\left\{-\frac{nt^2}{2\left(\lambda^\star\right)^2 k}\right\}.$$

Let $h(\cdot) = \phi_s(\cdot) - 1$ be an auxiliary function that is 1-Lipschitz and satisfies $h(0) = 0$. By classical symmetrization argument,

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}} \left|\hat{R}_{\phi_s}(f) - R_{\phi_s}(f)\right|\right] = \mathbb{E}\left[\sup_{f\in\mathcal{H}} \left|\hat{R}_h(f) - R_h(f)\right|\right] \leq 2\mathsf{Rad}_n\left(h\circ\mathcal{H}\right),$$

where

$$\mathsf{Rad}_n\left(h\circ\mathcal{H}\right) = \mathbb{E}\left[\sup_{f\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i h\left(Y_i f(X_i)\right)\right|\right]$$

and $\{\sigma_i\}_{i=1}^n$ are i.i.d. Rademacher variables ($\mathbb{P}[\sigma_i = +1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$). By contraction inequality (see Theorem 2.2 in [20]),

$$\mathsf{Rad}_n\left(h\circ\mathcal{H}\right) \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i Y_i f(X_i)\right|\right]$$

and hence

$$\begin{aligned}
\mathsf{Rad}_n\left(h \circ \mathcal{H}\right) &\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i Y_i \left\langle \mathcal{K}(X_i,\cdot), f\right\rangle_{\mathcal{H}_\mathcal{K}}\right|\right] \\
&\leq \sup_{\{(x_i,y_i)\}_{i=1}^n \subset \mathcal{X}\times\mathcal{Y}} \mathbb{E}\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i y_i \left\langle \mathcal{K}(x_i,\cdot), f\right\rangle_{\mathcal{H}_\mathcal{K}}\right|\right] \\
&\leq \frac{\lambda^\star}{n} \sup_{\{(x_i,y_i)\}_{i=1}^n \subset \mathcal{X}\times\mathcal{Y}} \sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{n}\sigma_i y_i \mathcal{K}(x_i,\cdot)\right\|^2_{\mathcal{H}_\mathcal{K}}\right]} \\
&= \frac{\lambda^\star}{n} \sup_{\{(x_i,y_i)\}_{i=1}^n \subset \mathcal{X}\times\mathcal{Y}} \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{K}\left(x_i,x_j\right) y_i y_j \mathbb{I}\left[i=j\right]} \\
&\leq \lambda^\star\sqrt{\frac{k}{n}}.
\end{aligned}$$

Aggregating all the results, we can obtain with probability at least $1-\delta$,

$$R(\hat{f}_n^{\phi_s}) - R(f^\star) \leq \inf_{f\in\mathcal{H}} R_{\phi_s}(f) - R^\star_{\phi_s} + 8\lambda^\star\sqrt{\frac{k}{n}} + 2\sqrt{\frac{2\left(\lambda^\star\right)^2 k}{n}\log\frac{2}{\delta}}.$$

References

[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[3] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[4] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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

[6] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *Proceedings of the 23rd international conference on Machine learning*, pages 201–208. ACM, 2006.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, and Manfred K Warmuth. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(3):551–585, 2006.

[10] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[11] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[12] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[13] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[14] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[15] Kevin Gimpel and Noah A Smith. Structured ramp loss minimization for machine translation. pages 221–231, 2012.

[16] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[17] Xiaolin Huang, Lei Shi, and Johan A. K. Suykens. Ramp loss linear programming support vector machine. *Journal of Machine Learning Research*, 15:2185–2211, 2014.

[18] Ling Jian, Fuhao Gao, Peng Ren, Yunquan Song, and Shihua Luo. A noise-resilient online learning algorithm for scene classification. *Remote Sensing*, 10(11):1836, 2018.

[19] Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004.

[20] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[21] Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine learning*, 78(3):287–304, 2010.

[22] H Brendan Mcmahan, Gary Holt, D Sculley, Michael Young, Dietmar Ebner, Julian Paul Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. pages 1222–1230, 2013.

[23] M. C. Du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In *In Advances in Neural Information Processing Systems*, pages 703–711, 2014.

[24] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

[25] 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.

[26] Yingjie Tian, Mahboubeh Mirzabagheri, Seyed Mojtaba Hosseini Bamakan, Huadong Wang, and Qiang Qu. Ramp loss one-class support vector machine; a robust and effective approach to anomaly detection problems. *Neurocomputing*, 310:223 – 235, 2018.

[27] Régis Vert and Jean Philippe Vert. Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research*, 7:817–854, 2006.

[28] Yingchao Xiao, Huangang Wang, and Wenli Xu. Ramp loss based robust one-class svm. *Pattern Recognition Letters*, 85:15–20, 2017.