Table 3: Classification results: error rate (standard deviation). The best performance figure in each dataset is denoted in **bold** typeface and the second best is denoted in ***bold italic***.

	Linear		Global Local		LSL-SP		LDKL		FaLK-SVM		RBF-SVM	
skin	8.900	(0.174)	0.249	(0.048)	12.481	(8.729)	1.858	(1.012)	**0.040**	(0.016)	***0.229***	(0.029)
winequality	33.685	(1.988)	***25.715***	(1.282)	30.878	(1.783)	36.795	(2.198)	28.706	(1.298)	**25.596**	(1.744)
census_income	43.972	(0.404)	***35.697***	(0.453)	**35.405**	(1.179)	47.229	(2.653)	–		45.843	(0.772)
twitter	6.964	(0.164)	***4.231***	(0.090)	8.370	(0.245)	15.557	(11.893)	**4.135**	(0.149)	9.109	(0.160)
a1a	***16.563***	(2.716)	**16.250**	(2.219)	20.438	(2.717)	17.063	(1.655)	18.125	(1.598)	16.500	(1.346)
breast-cancer	35.000	(4.402)	**3.529**	(1.683)	***3.677***	(2.110)	35.000	(4.402)	–		33.824	(4.313)
internet_ad	7.319	(1.502)	**2.638**	(1.003)	6.383	(1.118)	13.064	(3.601)	***3.362***	(0.997)	3.447	(0.772)

Table 4: Regression results: root mean squared loss (standard deviation). The best performance figure in each dataset is denoted in **bold** typeface and the second best is denoted in ***bold italic***.

	Linear	Global Local	RegTree	RBF-SVR
energy_heat	0.480 (0.047)	***0.101*** (0.014)	**0.050** (0.005)	0.219 (0.017)
energy_cool	0.501 (0.044)	**0.175** (0.018)	***0.200*** (0.018)	0.221 (0.026)
abalone	***0.687*** (0.024)	**0.659** (0.023)	0.727 (0.028)	0.713 (0.025)
kinematics	0.766 (0.019)	***0.634*** (0.022)	0.732 (0.031)	**0.347** (0.010)
puma8NH	0.793 (0.023)	***0.601*** (0.017)	0.612 (0.024)	**0.571** (0.020)
bank8FM	0.255 (0.012)	***0.218*** (0.009)	0.254 (0.008)	**0.202** (0.007)
communities	***0.586*** (0.049)	**0.578** (0.040)	0.653 (0.060)	0.618 (0.053)

For our models, we used logistic functions for loss functions. The max iteration number was set as 1000, and the algorithm stopped early when the gap in the empirical loss from the previous iteration became lower than 10^{-9} in 10 consecutive iterations. Hyperparameters[9] were optimized through 10-fold cross validation. We fixed the number of regions to 10 in LSL-SP, tree-depth to 3 in LDKL, and neighborhood size to 100 in FaLK-SVM.

Table 3 summarizes the classification errors. We observed 1) Global/Local consistently performed well and achieved the best error rates foir four datasets out of seven. 2) LSL-SP performed well for census_income and breast-cancer, but did significantly worse than Linear for skin, twitter, and a1a. Similarly, LDKL performed worse than Linear for census_income, twitter, a1a and internet_ad. This arose partly because of over fitting and partly because of bad local minima. Particularly noteworthy is that the standard deviations in LDKL were much larger than in the others, and the initialization issue would seem to become significant in practice. 3) FaLK-SVM performed well in most cases, but its computational cost was significantly higher than that of others, and it was unable to obtain results for census_income and internet_ad (we stopped the algorithm after 24 hours running).

5.2.2 Regression

For regression, we compared Global/Local with Linear, regression tree[10] by CART (RegTree) [1], and epsilon-SVR with RBF kernel[11]. Target variables were standardized so that their mean was 0 and their variance was 1. Performance was evaluated using the root mean squared loss in the test data. Tree-depth of RegTree and ϵ in RBF-SVR were determined by means of 10-fold cross validation. Other experimental settings were the same as those used in the classification tasks.

Table 4 summarizes RMSE values. In classification tasks, Global/Local consistently performed well. For the kinematics, RBF-SVR performed much better than Global/Local, but Global/Local was better than Linear and RegTree in many other datasets.

6 Conclusion

We have proposed here a novel convex formulation of region-specific linear models that we refer to as partition-wise linear models. Our approach simultaneously optimizes regions and predictors using sparsity-inducing structured penalties. For the purpose of efficiently solving the optimization problem, we have derived an efficient algorithm based on the decomposition of proximal maps. Thanks to its convexity, our method is free from initialization dependency, and a generalization error bound can be derived. Empirical results demonstrate the superiority of partition-wise linear models over other region-specific and locally linear models.

Acknowledgments
The majority of the work was done during the internship of the first author at the NEC central research laboratories.

[9] λ^1, λ^2_p in Global/Local,λ^1 in Linear, $\lambda_W, \lambda_\theta, \lambda_{\theta'}, \sigma$ in LDKL, C in FaLK-SVM, and C, γ in RBF-SVM.
[10] We used a scikit-learn package. http://scikit-learn.org/
[11] We used a libsvm package.