Launch and Iterate: Reducing Prediction Churn Appendix

Q. Cormier
ENS Lyon
15 parvis René Descartes
Lyon, France
`quentin.cormier@ens-lyon.fr`

M. Milani Fard, K. Canini, M. R. Gupta
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
`{mmilanifard,canini,mayagupta}@google.com`

This appendix includes proofs of the theorems presented in the paper and details about the experimental design.

A Proof of Theorem 1

Proof. Let $z = (x, y)$ be a fixed testing sample. We define the function $G_z : (\mathcal{X} \times \mathcal{Y})^{2m} \to \mathbb{R}$ as:

$$G_z(T, T') = \ell_\gamma(f_T(x), y) - \ell_\gamma(f_{T'}(x), y). \tag{1}$$

For any $i \in \{1, .., m\}$, if we re-sample the ith element in T to get T^i, using the β-stability of the learning algorithm and Lipschitz continuity of ℓ_γ we get:

$$|G_z(T, T') - G_z(T^i, T')| \leq \frac{1}{\gamma} |f_T(x) - f_{T^i}(x)| \leq \frac{\beta}{\gamma}. \tag{2}$$

The same inequality holds for $|G_z(T, T') - G_z(T, {T'}^i)|$. We have $\mathbb{E}_{T,T' \sim D^m}[G_z(T, T')] = 0$, and thus can apply the McDiarmid inequality to get:

$$\Pr_{T,T' \sim D^m}[|G_z(T, T')| > \epsilon] \leq 2e^{-\frac{\epsilon^2 \gamma^2}{m\beta^2}}. \tag{3}$$

Integrating the above gives us the bound over the expectation:

$$\mathop{\mathbb{E}}_{T,T' \sim \mathcal{D}^m}[|G_z(T, T')|] \leq \int_0^\infty \Pr_{T,T' \sim \mathcal{D}^m}[|G_z(T, T')| > \epsilon] d\epsilon \leq \frac{\beta\sqrt{\pi m}}{\gamma}. \tag{4}$$

The above inequality holds for any fixed z and thus holds for the expectation:

$$\begin{aligned}
\mathop{\mathbb{E}}_{T,T' \sim \mathcal{D}^m}[C_\gamma(f_1, f_2)] &= \mathop{\mathbb{E}}_{T,T' \sim \mathcal{D}^m}\left[\mathop{\mathbb{E}}_{Z \sim \mathcal{D}}[|G_Z(T, T')|]\right] & (5)\\
&= \mathop{\mathbb{E}}_{Z \sim \mathcal{D}}\left[\mathop{\mathbb{E}}_{T,T' \sim \mathcal{D}^m}[|G_Z(T, T')|]\right] & (6)\\
&\leq \frac{\beta\sqrt{\pi m}}{\gamma}. & (7)
\end{aligned}$$

□

B Proof of Theorem 2

Proof. We again use the McDiarmid inequality, on the function $H : (\mathcal{X} \times \mathcal{Y})^{2m} \to \mathbb{R}$ defined as:

$$H(T, T') = C_\gamma(f_T, f_{T'}) = \mathop{\mathbb{E}}_{(X,Y) \sim \mathcal{D}}[|\ell_\gamma(f_T(X), Y) - \ell_\gamma(f_{T'}(X), Y)|]. \tag{8}$$

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

For any $i \in \{1, .., m\}$, if we re-sample the ith element in T to get T^i, using the β-stability of the learning algorithm and Lipschitz continuity of ℓ_γ we get:

$$\begin{aligned} |H(T,T') - H(T^i,T')| &\leq \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} \Big[||\ell_\gamma(f_T(X),Y) - \ell_\gamma(f_{T'}(X),Y)| - && (9) \\ & \qquad |\ell_\gamma(f_{T^i}(X),Y) - \ell_\gamma(f_{T'}(X),Y)|| \Big] && (10) \\ &\leq \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} [|\ell_\gamma(f_T(X),Y) - \ell_\gamma(f_{T^i}(X),Y)|] && (11) \\ &\leq \frac{\beta}{\gamma}, && (12) \end{aligned}$$

where line (11) is by reverse triangular inequality. Same bound similarly holds for replacing the ith element in T': $|H(T,T') - H(T,{T'}^i)| \leq \beta/\gamma$. Applying McDiarmid inequality and using the bound on the expectation of H from Theorem 1 completes the proof:

$$\begin{aligned} \Pr_{T,T'\sim\mathcal{D}^m} \left\{ C_\gamma(f_T, f_{T'}) > \epsilon + \frac{\sqrt{\pi m}\beta}{\gamma} \right\} &\leq \Pr_{T,T'\sim\mathcal{D}^m} \left\{ H(T,T') > \epsilon + \mathop{\mathbb{E}}_{T,T'\sim\mathcal{D}^m}[H(T,T')] \right\} \\ &\leq e^{-\frac{\epsilon^2\gamma^2}{m\beta^2}}. && (13) \end{aligned}$$

□

C Proof of Theorem 3

The proof partly follows Lemma 21 from [1] and Theorem 4 from [2]. Define:

$$\begin{aligned} \ell_j(g) &= (g(x_j) - y_j)^2 && (14) \\ \hat{R}_T(g) &= \frac{1}{m}\sum_{j=1}^{m} \ell_j(g) && (15) \\ \hat{R}_T^{\setminus i}(g) &= \frac{1}{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} \ell_j(g) && (16) \\ R_T(g) &= \hat{R}_T(g) + \lambda\|g\|_k^2 && (17) \\ R_T^{\setminus i}(g) &= \hat{R}_T^{\setminus i}(g) + \lambda\|g\|_k^2. && (18) \end{aligned}$$

By the assumption of the theorem, f_T is the minimizer of R_T. Let $f_T^{\setminus i}$ be the minimizer of $R_T^{\setminus i}$.

Lemma 1. *With the assumptions of Theorem 3, we have for all i:*

$$\forall x : (f_T(x) - f_T^{\setminus i}(x))^2 \leq \frac{\kappa^4}{\lambda^2 m^2}(f_T(x_i) - y_i)^2. \tag{19}$$

Proof of Lemma 1. To simply the notation, we drop the T subscript throughout the proof of this lemma. Let $d_\phi(f,g)$ be the functional Bregman divergence [3]:

$$d_\phi(f,g) = \phi(f) - \phi(g) - \nabla\phi(g; f-g), \tag{20}$$

where $\nabla\phi(g;.)$ is the Fréchet derivative of ϕ at g. Since f and $f^{\setminus i}$ are minimizers of R and $R^{\setminus i}$ respectively, we have: $\nabla R(f;.) = 0$ and $\nabla R^{\setminus i}(f^{\setminus i};.) = 0$. We thus have:

$$\begin{aligned} d_R(f^{\setminus i}, f) + d_{R^{\setminus i}}(f, f^{\setminus i}) &= R(f^{\setminus i}) - R(f) + R^{\setminus i}(f) - R^{\setminus i}(f^{\setminus i}) && (21) \\ &= \frac{1}{m}\ell_i(f^{\setminus i}) - \frac{1}{m}\ell_i(f), && (22) \end{aligned}$$

where the last line follows by the definition of R and $R^{\setminus i}$. By non-negativity and additivity of divergence ($d_{A+B} = d_A + d_B$) we have:

$$
\begin{aligned}
0 &\leq d_{\hat{R}^{\setminus i}}(f, f^{\setminus i}) + d_{\hat{R}^{\setminus i}}(f^{\setminus i}, f) && (23)\\
&= -\lambda d_{\|.\|_k^2}(f, f^{\setminus i}) - \lambda d_{\|.\|_k^2}(f^{\setminus i}, f) + d_{R^{\setminus i}}(f, f^{\setminus i}) + d_{R^{\setminus i}}(f^{\setminus i}, f) && (24)\\
&= -\lambda d_{\|.\|_k^2}(f, f^{\setminus i}) - \lambda d_{\|.\|_k^2}(f^{\setminus i}, f) + d_{R^{\setminus i}}(f, f^{\setminus i}) + d_{R}(f^{\setminus i}, f) - \frac{1}{m} d_{\ell_i}(f^{\setminus i}, f) && (25)\\
&= -\lambda d_{\|.\|_k^2}(f, f^{\setminus i}) - \lambda d_{\|.\|_k^2}(f^{\setminus i}, f) + \frac{1}{m}\ell_i(f^{\setminus i}) - \frac{1}{m}\ell_i(f) - \frac{1}{m} d_{\ell_i}(f^{\setminus i}, f) && (26)\\
&= -\lambda d_{\|.\|_k^2}(f, f^{\setminus i}) - \lambda d_{\|.\|_k^2}(f^{\setminus i}, f) + \frac{1}{m}\nabla\ell_i(f; f^{\setminus i} - f), && (27)
\end{aligned}
$$

where line (26) is by the derivation in line (22), and line (27) is by the definition of the Bregman divergence. In the RKHS space, we have $d_{\|.\|_k^2}(g, g') = \|g - g'\|_k^2$, and by assumption of Theorem 3 we have $\forall x : |g(x)| \leq \kappa \|g\|_k$. Substituting the Fréchet derivative in the above inequality, we get:

$$
\begin{aligned}
\|f - f^{\setminus i}\|_k^2 &\leq \frac{1}{\lambda m}(f^{\setminus i}(x) - f(x))(f(x_i) - y_i) && (28)\\
&\leq \frac{\kappa}{\lambda m}\|f^{\setminus i} - f\|_k (f(x_i) - y_i). && (29)
\end{aligned}
$$

Cancelling the sides and squaring both sides, we get for all x:

$$
\begin{aligned}
(f(x) - f^{\setminus i}(x))^2 &\leq \kappa^2 \|f - f^{\setminus i}\|_k^2 && (30)\\
&\leq \frac{\kappa^4}{\lambda^2 m^2}(f(x_i) - y_i)^2. && (31)
\end{aligned}
$$

□

Proof of Theorem 3. Let $V = \ell_\gamma(f_T(X), Y) - \ell_\gamma(f_{T'}(X), Y)$. Define V_i, $1 \leq i \leq 2m$ as:

$$
V_i = \begin{cases} \ell_\gamma(f_T^{\setminus i}(X), Y) - \ell_\gamma(f_{T'}(X), Y) & \text{if } i \leq m \\ \ell_\gamma(f_T(X), Y) - \ell_\gamma(f_{T'}^{\setminus (i-m)}(X), Y) & \text{if } i > m \end{cases} \tag{32}
$$

It is easy to see that $\mathbb{E}_{T,T' \sim \mathcal{D}^m}[V] = 0$. Using the concentration inequality of Theorem 6 from [4] on V and V_i, the symmetry of the training algorithm, and the symmetry of V on T and T' we get:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\substack{T,T' \sim \mathcal{D}^m \\ (X,) \sim \mathcal{D}}}[(\ell_\gamma(f_T(X), Y) - \ell_\gamma(f_{T'}(X), Y))^2] &= \mathop{\mathbb{V}\mathrm{ar}}_{\substack{T,T' \sim \mathcal{D}^m \\ (X,Y) \sim \mathcal{D}}}[V] && (33)\\
&\leq \sum_{i=1}^{2m} \mathop{\mathbb{E}}_{\substack{T,T' \sim \mathcal{D}^m \\ (X,Y) \sim \mathcal{D}}}[(V - V_i)^2] && (34)\\
&= 2 \mathop{\mathbb{E}}_{\substack{T,T' \sim \mathcal{D}^m \\ (X,Y) \sim \mathcal{D}}}\left[\sum_{i=1}^{m}(V - V_i)^2\right] && (35)\\
&= \frac{2}{\gamma^2} \mathop{\mathbb{E}}_{\substack{T,T' \sim \mathcal{D}^m \\ (X,Y) \sim \mathcal{D}}}\left[\sum_{i=1}^{m}(f_T(X) - f_T^{\setminus i}(X))^2\right], && (36)
\end{aligned}
$$

where line (36) is by Lipschitz continuity of ℓ_γ. Applying Lemma 1 to RHS completes the proof. □

D Further Experimental Details

Table 1 includes further details on the datasets used for experiments presented in the paper.

Table 1: Full details of the datasets used in the experimental analysis.

	Nomao [5]	News Popularity [6]	Twitter Buzz [7]
# Features	89 continuous 31 nominal some missing values	61 features no missing values	77 features evolution of 11 primatry features through time no missing values
# Samples	34,465	39,797	Sub-sampled 46,902
Goal	predict if two business entities are the same	predict if a news will be shared more than 1400 times	predict if a tweet is going to be popular
T_A	4000 samples drop first 5 features	8000 samples drop the 3 features: *self_reference_min* *self_reference_max* *self_reference_avg*	4000 samples drop last 7 features
T_B	5000 samples all the features	10000 samples all the features	5000 samples all the features
Validation Set	1000 samples	1000 samples	1000 samples
Testing Set	28465 samples	28797 samples	45402 samples

We optimized the hyper-parameters of each algorithm for each datasets on the validation set. Details of the chosen hyper-parameters for each algorithm is included in Table 2. The names of the parameters match the names used in Scikit-Learn [8].

Table 2: We summarize here the regularization parameters used to train the models. These parameters have been selected using a validation set of 1000 samples.

	Ridge α	RFT-Regression min_weight_fraction_leaf	SVM C	Adaboost learning_rate	LinearSVR C
Nomao	0.02	0.0001	10	1.5	0.5
News	2	0.01	1.5	5	10
Twitter-Buzz	1	0.002	50	1.0	75

Full results for all experiments are included in Table 3. We have included further results on linear SVM and AdaBoost (boosted stumps). However, note that there is a regression in accuracy between the two versions of the model for the baseline algorithm. We believe that that our hyper-parameter optimization did not find a good solution for these algorithms (likely resulting in over-fitting), or that we could not effectively use the implementation in Scikit-Learn [8].

Table 3: Experiment results on 3 domains with 5 different training algorithms for a single step RCP and the MCMC methods. For the MCMC experiment, we report the numbers with the standard deviation over the 40 runs of the chain.

			Baseline No RCP, No Chain	RCP $\alpha = 0.5, \epsilon = 0.5$	MCMC, $k = 30$ $\alpha = 0.5, \epsilon = 0.5$	MCMC, $k = 30$ $\alpha = 0.7, \epsilon = 0.1$
Nomao	Ridge	WLR	1.24	1.40	1.31	1.60
		p_{win}	26.5	49.2	36.5	73.9
		C_r	1.00	0.54	0.54 ± 0.06	0.32 ± 0.05
		Acc V_1 / V_2	93.1 / 93.4	93.1 / 93.4	93.2 ± 0.1 / 93.4 ± 0.1	93.0 ± 0.3 / 93.2 ± 0.2
	RF	WLR	1.02	1.13	1.09	1.12
		p_{win}	5.6	13.4	9.8	13.1
		C_r	1.00	0.83	0.83 ± 0.05	0.59 ± 0.05
		Acc V_1 / V_2	94.8 / 94.8	94.8 / 95.0	94.9 ± 0.2 / 95.0 ± 0.2	94.7 ± 0.2 / 94.8 ± 0.2
	AdaBoost	WLR	0.79	0.79	0.00	0.00
		p_{win}	0.2	0.2	0.0	0.0
		C_r	1.00	1.00	0.01 ± 0.06	0.00 ± 0.00
		Acc V_1 / V_2	85.7 / 85.3	85.7 / 85.3	75.7 ± 2.4 / 75.6 ± 2.2	77.4 ± 6.2 / 77.4 ± 6.2
	LinSVM	WLR	0.64	0.89	0.90	2.60
		p_{win}	0.0	1.2	1.3	99.8
		C_r	1.00	0.75	0.76 ± 0.02	0.22 ± 0.02
		Acc V_1 / V_2	90.1 / 86.2	90.1 / 89.3	90.1 ± 0.4 / 89.4 ± 0.3	90.1 ± 0.5 / 91.9 ± 0.5
	SVM	WLR	1.70	2.51	2.32	2.08
		p_{win}	82.5	99.7	99.2	97.1
		C_r	1.00	0.75	0.69 ± 0.06	0.54 ± 0.03
		Acc V_1 / V_2	94.6 / 95.1	94.6 / 95.2	94.8 ± 0.2 / 95.3 ± 0.1	94.9 ± 0.2 / 95.2 ± 0.1
News	Ridge	WLR	0.95	0.94	1.04	0.97
		p_{win}	2.5	2.4	6.7	3.4
		C_r	1.00	0.75	0.78 ± 0.04	0.42 ± 0.06
		Acc V_1 / V_2	65.1 / 65.0	65.1 / 65.0	65.0 ± 0.1 / 65.1 ± 0.1	64.7 ± 0.2 / 64.7 ± 0.2
	RF	WLR	1.07	1.02	1.10	1.24
		p_{win}	8.5	5.7	10.8	26.6
		C_r	1.00	0.69	0.67 ± 0.04	0.04 ± 0.04
		Acc V_1 / V_2	64.5 / 65.1	64.5 / 64.7	64.3 ± 0.3 / 64.8 ± 0.2	63.0 ± 0.4 / 63.0 ± 0.4
	AdaBoost	WLR	0.72	0.72	0.81	0.00
		p_{win}	0.0	0.0	0.3	0.0
		C_r	1.00	1.00	7.88 ± 12.07	0.03 ± 0.06
		Acc V_1 / V_2	59.3 / 59.2	59.3 / 59.2	59.4 ± 0.2 / 59.2 ± 0.0	58.7 ± 1.1 / 58.7 ± 1.1
	LinSVM	WLR	0.81	1.24	1.03	1.02
		p_{win}	0.3	26.4	6.1	5.3
		C_r	1.00	0.90	1.10 ± 0.19	1.12 ± 0.26
		Acc V_1 / V_2	63.3 / 62.5	63.3 / 64.1	63.5 ± 0.5 / 63.6 ± 0.5	63.0 ± 0.8 / 63.1 ± 0.7
	SVM	WLR	1.17	1.26	1.24	1.25
		p_{win}	18.4	29.4	26.1	28.0
		C_r	1.00	0.77	0.86 ± 0.02	0.61 ± 0.02
		Acc V_1 / V_2	64.9 / 65.4	64.9 / 65.4	64.8 ± 0.1 / 65.4 ± 0.1	64.7 ± 0.2 / 65.1 ± 0.1
Twitter Buzz	Ridge	WLR	1.71	3.54	1.53	1.58
		p_{win}	83.1	100.0	66.4	71.9
		C_r	1.00	0.85	0.65 ± 0.05	0.44 ± 0.04
		Acc V_1 / V_2	89.7 / 89.9	89.7 / 90.0	90.1 ± 0.1 / 90.2 ± 0.1	89.7 ± 0.1 / 89.7 ± 0.1
	RF	WLR	1.35	1.15	1.15	1.03
		p_{win}	41.5	16.1	15.9	6.0
		C_r	1.00	0.86	0.77 ± 0.07	0.42 ± 0.10
		Acc V_1 / V_2	96.2 / 96.4	96.2 / 96.3	96.3 ± 0.1 / 96.3 ± 0.1	96.2 ± 0.1 / 96.2 ± 0.1
	AdaBoost	WLR	0.93	0.90	1.13	1.17
		p_{win}	1.8	1.2	13.3	18.4
		C_r	1.00	1.03	0.80 ± 0.18	0.22 ± 0.07
		Acc V_1 / V_2	95.0 / 95.0	95.0 / 95.0	94.2 ± 0.4 / 94.2 ± 0.4	95.5 ± 0.3 / 95.5 ± 0.3
	LinSVM	WLR	0.22	2.66	3.71	3.82
		p_{win}	0.0	99.9	100.0	100.0
		C_r	1.00	0.52	0.61 ± 0.45	0.41 ± 0.22
		Acc V_1 / V_2	94.8 / 91.2	94.8 / 96.2	92.2 ± 2.7 / 92.7 ± 2.5	93.0 ± 2.0 / 93.2 ± 2.0
	SVM	WLR	1.35	1.77	1.55	1.33
		p_{win}	42.2	86.6	68.4	39.3
		C_r	1.00	0.70	0.70 ± 0.03	0.50 ± 0.03
		Acc V_1 / V_2	96.0 / 96.1	96.0 / 96.1	96.1 ± 0.1 / 96.2 ± 0.1	96.1 ± 0.1 / 96.2 ± 0.1

References

[1] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

[2] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, volume 13, page 196. MIT Press, 2001.

[3] B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *Information Theory, IEEE Transactions on*, 54(11):5130–5139, 2008.

[4] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*, pages 208–240. Springer, 2004.

[5] L. Candillier and V. Lemaire. Design and analysis of the Nomao challenge active learning in the real-world. In *Proceedings of the ALRA: Active Learning in Real-world Applications, Workshop ECML-PKDD*, 2012.

[6] K. Fernandes, P. Vinagre, and P. Cortez. *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings*, chapter A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News, pages 535–546. Springer International Publishing, Cham, 2015.

[7] F. Kawala, E. Gaussier, A. Douzal-Chouakria, and E. Diemert. Apprentissage d'ordonnancement et influence de l'ambiguïté pour la prédiction d'activité sur les réseaux sociaux. In *Coria'2014*, pages 1–15, Nancy, France, France, March 2014.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.