Supplementary Material for *"Large-scale Online Kernel Learning with Random Feature Reparameterization"*

Tu Dinh Nguyen†, Trung Le†, Hung Bui‡, Dinh Phung†
‡Adobe Research, Adobe Systems, Inc.
†Center for Pattern Recognition and Data Analytics, Deakin University, Australia
{tu.nguyen, trung.l, dinh.phung}@deakin.edu.au, hubui@adobe.com

This document presents supplementary material to complement the manuscript entitled "*Reparameterized Random Feature for Large-scale Online Kernel Learning*", accepted at the 26^{th} International Joint Conference on Artificial Intelligence (IJCAI), 2017. We first provide definitions of loss functions and their gradients w.r.t parameters. We then present full derivations of some equations in theoretical analysis, and add more details on gradient learning of our proposed model.

1 LOSS FUNCTIONS

We use 5 loss functions in our proposed model, namely Hinge, logistic, ℓ_2, ℓ_1 and ε-insensitive. Let $\mathbb{I}\{s\}$ denote the indicator function that returns 1 if the logical statement s is true, and 0 otherwise. These losses and their gradients w.r.t parameters $\mathbf{w}$ are listed in Table 1.

Loss	$\ell\left(\mathrm{y}, \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)$	$\nabla_{\mathbf{w}} \ell\left(\mathrm{y}, \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)$
Hinge	$\max\left(0, 1 - \mathrm{y}\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)$	$-\mathbb{I}\left\{\mathrm{y}\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) \leq 1\right\} \mathrm{y}\boldsymbol{\phi}(\mathbf{x})$
Logistic	$\log\left[1 + \exp\left(-\mathrm{y}\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)\right]$	$-\mathrm{y}\boldsymbol{\phi}(\mathbf{x}) \frac{\exp\left(-\mathrm{y}\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)}{\exp(-\mathrm{y}\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}))+1}$
ℓ_2	$\frac{1}{2}\left(\mathrm{y} - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)^2$	$\left(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathrm{y}\right) \boldsymbol{\phi}(\mathbf{x})$
ℓ_1	$\left\lvert \mathrm{y} - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) \right\rvert$	$\mathrm{sign}\left(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathrm{y}\right) \boldsymbol{\phi}(\mathbf{x})$
ε-insensitive	$\max\left\{0, \left\lvert \mathrm{y} - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) \right\rvert - \varepsilon\right\}$	$\mathbb{I}\left\{\left\lvert \mathrm{y} - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) \right\rvert > \varepsilon\right\}$ $\times \mathrm{sign}\left(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathrm{y}\right) \boldsymbol{\phi}(\mathbf{x})$

Table 1: Typical loss functions and their gradients.

For *multiclass classification*, we follow the work of Crammer and Singer (2001) in which the objective function reads:

$$\mathcal{J}(\mathbf{w}) \triangleq \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \mathbb{E}_{\mathbb{P}}\left[\ell\left(\mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x})\right)\right]$$

where $\mathrm{z} = \mathrm{argmax}_{k \neq \mathrm{y}} \mathbf{w}_k^\top \boldsymbol{\phi}(\mathbf{x})$ and the loss function can be either Hinge loss or logistic loss as listed in Table 2.

Loss	Hinge	Logistic
$\ell\left(\mathrm{y}, \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)$	$\max\left(0, 1 - \mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x}) + \mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x})\right)$	$\log\left[1 + \exp\left(\mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x})\right)\right]$
$\nabla_{\mathbf{w}_{\mathrm{y}}} \ell\left(\mathrm{y}, \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)$	$-\mathbb{I}\left\{\mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x}) \leq 1\right\} \boldsymbol{\phi}(\mathbf{x})$	$-\boldsymbol{\phi}(\mathbf{x}) \frac{\exp\left(\mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x})\right)}{\exp\left(\mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x})\right)+1}$
$\nabla_{\mathbf{w}_{\mathrm{z}}} \ell\left(\mathrm{y}, \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})\right)$	$\mathbb{I}\left\{\mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x}) \leq 1\right\} \boldsymbol{\phi}(\mathbf{x})$	$\boldsymbol{\phi}(\mathbf{x}) \frac{\exp\left(\mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x})\right)}{\exp\left(\mathbf{w}_{\mathrm{z}}^\top \boldsymbol{\phi}(\mathbf{x}) - \mathbf{w}_{\mathrm{y}}^\top \boldsymbol{\phi}(\mathbf{x})\right)+1}$

Table 2: Hinge loss, logistic loss and their gradients for multiclass classification.

2 REPARAMETERIZED RANDOM FEATURES

In this section, we first roughly demonstrate the fast convergence of the upper bound of approximation error probability proven in Lemma 2 in the manuscript. Fig. 1 illustrates that the bound exponentially converges to zero when the number of random features passes a certain point for each precision ε.

We now present full derivations of some equations in Section 3.1 in the manuscript.

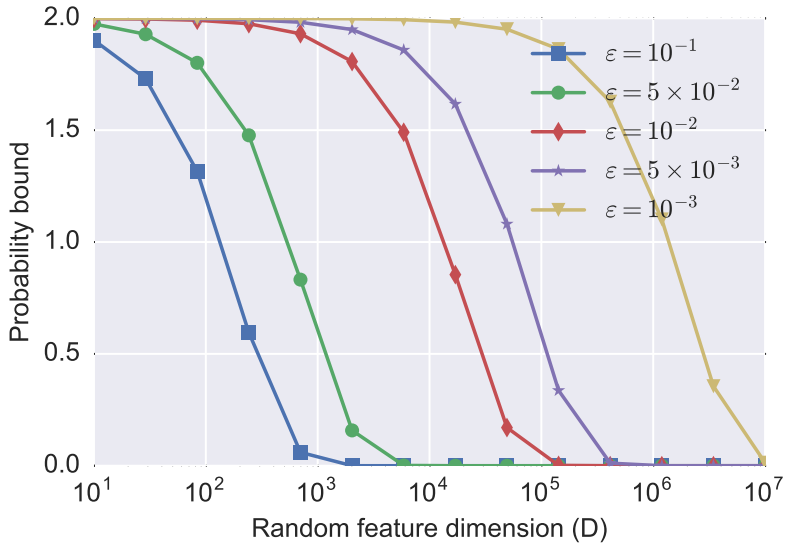

Figure 1: The convergence of the upper bound of approximation error probability.

- Eq (6) in the manuscript:

$$
\begin{aligned}
\mathbb{E}\left[L_g^2\right] &= \mathbb{E}\left[\left\|\nabla g\left(\mathbf{u}^\star\right)\right\|^2\right] = \mathbb{E}\left[\left\|\nabla \tilde{k}_{\boldsymbol{\epsilon}}\left(\mathbf{u}^\star\right) - \nabla k\left(\mathbf{u}^\star\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla \tilde{k}_{\boldsymbol{\epsilon}}\left(\mathbf{u}^\star\right)\right\|^2\right] - \left\|\nabla k\left(\mathbf{u}^\star\right)\right\|^2 \leq \mathbb{E}\left[\left\|\nabla \tilde{k}_{\boldsymbol{\epsilon}}\left(\mathbf{u}^\star\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{\mathrm{D}}\sum_{d=1}^{\mathrm{D}} \sin\left(\left(\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}_d\right)^\top \mathbf{u}^\star\right)\left(\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}_d\right)^\top\right\|^2\right] \\
&\leq \frac{1}{\mathrm{D}^2}\mathbb{E}\left[\left(\sum_{d=1}^{\mathrm{D}}\left\|\sin\left(\left(\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}_d\right)^\top \mathbf{u}^\star\right)\left(\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}_d\right)^\top\right\|\right)^2\right] \\
&\leq \frac{1}{\mathrm{D}^2}\mathbb{E}\left[\left(\sum_{d=1}^{\mathrm{D}}\left\|\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}_d\right\|\right)^2\right] \qquad (1) \\
&\leq \frac{1}{\mathrm{D}^2}\mathbb{E}\left[\mathrm{D}\sum_{d=1}^{\mathrm{D}}\left\|\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}_d\right\|^2\right] \qquad (2) \\
&= \frac{1}{\mathrm{D}}\sum_{d=1}^{\mathrm{D}}\mathbb{E}\left[\left\|\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}_d\right\|^2\right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}\right\|^2\right] = \left\|\boldsymbol{\sigma}\right\|^2 \qquad (3)
\end{aligned}
$$

Here we note that to go from Eq. (1) to Eq. (2), we have used Schwartz-Cauchy inequality $\left(\sum_{d=1}^{\mathrm{D}} a_d\right)^2 \leq \mathrm{D}\sum_{d=1}^{\mathrm{D}} a_d^2$ and to obtain Eq. (3) we derive as:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\operatorname{diag}\left(\boldsymbol{\sigma}\right)\boldsymbol{\epsilon}\right\|^2\right] &= \sum_{n=1}^{\mathrm{N}}\sigma_n^2\mathbb{E}_{\epsilon_n\sim\mathcal{N}(0,1)}\left[\epsilon_n^2\right] \\
&= \sum_{n=1}^{\mathrm{N}}\sigma_n^2 = \left\|\boldsymbol{\sigma}\right\|^2
\end{aligned}
$$

where ϵ_n^2 follows Chi square distribution with 1 degree of freedom, and hence $\mathbb{E}\left[\epsilon_n^2\right] = 1$.

- Eq (7) in the manuscript:

$$
\begin{aligned}
&2\sqrt{\tau_1\tau_2\left(\frac{\tau_1}{\tau_2}\right)^{\frac{-\mathrm{N}+2}{\mathrm{N}+2}}} = 2\tau_1^{\frac{2}{\mathrm{N}+2}}\tau_2^{\frac{\mathrm{N}}{\mathrm{N}+2}} \\
&= 2\times 4^{\frac{2\mathrm{N}+1}{\mathrm{N}+2}}\left(\frac{\operatorname{diam}(\mathcal{X})\left\|\boldsymbol{\sigma}\right\|^2}{\varepsilon^2}\right)^{\frac{\mathrm{N}}{\mathrm{N}+2}}\exp\left(\frac{-\mathrm{D}\varepsilon^2}{\mathrm{N}+2}\right) \\
&\leq 2^5\exp\left(-\frac{\mathrm{D}\varepsilon^2}{\mathrm{N}+2}+\frac{\mathrm{N}}{\mathrm{N}+2}\log\left(\frac{\operatorname{diam}(\mathcal{X})\left\|\boldsymbol{\sigma}\right\|^2}{\varepsilon^2}\right)\right) \\
&\leq 2^5\exp\left(-\frac{\mathrm{D}\varepsilon^2}{\mathrm{N}+2}+\log\left(\frac{\operatorname{diam}(\mathcal{X})\left\|\boldsymbol{\sigma}\right\|^2}{\varepsilon^2}\right)\right) \\
&= 2^5\left(\frac{\operatorname{diam}(\mathcal{X})\left\|\boldsymbol{\sigma}\right\|^2}{\varepsilon^2}\right)\exp\left(-\frac{\mathrm{D}\varepsilon^2}{\mathrm{N}+2}\right)
\end{aligned}
$$

Next we provide details on derivations of gradients to learn the kernel parameters in Section 3.2 in the manuscript. Recall that the gradient of loss function w.r.t to kernel width parameters $\boldsymbol{\sigma}$ in Eq. (10) in the manuscript is computed by:

$$
\begin{aligned}
\nabla_{\boldsymbol{\sigma}}\ell\left(\mathrm{y},\mathbf{w}^\top\tilde{\boldsymbol{\phi}}\left(\mathbf{x}\right)\right) &= \sum_{d=1}^{\mathrm{D}}\left(\frac{\partial\left[\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\right]_d\left(\mathbf{x}\right)}{\partial\boldsymbol{\omega}_d}\right)^\top\frac{\partial\ell\left(\mathrm{y},\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)}{\partial\left[\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\right]_d\left(\mathbf{x}\right)}\odot\frac{\partial\boldsymbol{\omega}_d}{\partial\boldsymbol{\sigma}} \\
&= \sum_{d=1}^{\mathrm{D}}\left[-\mathbf{x}\sin\left(\boldsymbol{\omega}_d^\top\mathbf{x}\right),\mathbf{x}\cos\left(\boldsymbol{\omega}_d^\top\mathbf{x}\right)\right]^\top \\
&\quad\times\left(\frac{\partial\ell\left(\mathrm{y},\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)}{\partial\left[\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\right]_d\left(\mathbf{x}\right)}\right)^\top\odot\boldsymbol{\epsilon}_d
\end{aligned}
$$

Similarity to those w.r.t $\mathbf{w}$, the derivative of loss function ℓ w.r.t $\left[\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right]_d$ can be computed as shown in Table 3.

Loss	$\ell\left(\mathrm{y},\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)$	$\nabla_{\left[\hat{\mathbf{z}}_{\boldsymbol{\sigma}}(\mathbf{x})\right]_d}\ell\left(\mathrm{y},\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)$
Hinge	$\max\left(0,1-\mathrm{y}\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)$	$-\mathbb{I}\left\{\mathrm{y}\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\leq 1\right\}\mathrm{y}\mathbf{w}$
Logistic	$\log\left[1+\exp\left(-\mathrm{y}\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)\right]$	$-\mathrm{y}\mathbf{w}\frac{\exp\left(-\mathrm{y}\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}(\mathbf{x})\right)}{\exp(-\mathrm{y}\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}(\mathbf{x}))+1}$
ℓ_2	$\frac{1}{2}\left(\mathrm{y}-\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)^2$	$\left(\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)-\mathrm{y}\right)\mathbf{w}$
ℓ_1	$\left\lvert\mathrm{y}-\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right\rvert$	$\operatorname{sign}\left(\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)-\mathrm{y}\right)\mathbf{w}$
ε-insensitive	$\max\left\{0,\left\lvert\mathrm{y}-\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right\rvert-\varepsilon\right\}$	$\mathbb{I}\left\{\left\lvert\mathrm{y}-\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right\rvert>\varepsilon\right\}$ $\times\operatorname{sign}\left(\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)-\mathrm{y}\right)\mathbf{w}$

Table 3: Typical loss functions and their gradients w.r.t $\left[\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right]_d$.

As the variance $\boldsymbol{\sigma}$ is constrained to be positive, we can learn the log-variance $\boldsymbol{\gamma}=\log\left(\boldsymbol{\sigma}\right)$ instead, hence naturally induce a positive variance. The gradient of loss function w.r.t to $\boldsymbol{\gamma}$ can be computed as:

$$
\begin{aligned}
\nabla_{\boldsymbol{\gamma}}\ell\left(\mathrm{y},\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right) &= \nabla_{\boldsymbol{\sigma}}\ell\left(\mathrm{y},\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)\odot\nabla_{\boldsymbol{\gamma}}\boldsymbol{\sigma} \\
&= \nabla_{\boldsymbol{\sigma}}\ell\left(\mathrm{y},\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)\odot\nabla_{\boldsymbol{\gamma}}e^{\boldsymbol{\gamma}} \\
&= \nabla_{\boldsymbol{\sigma}}\ell\left(\mathrm{y},\mathbf{w}^\top\hat{\mathbf{z}}_{\boldsymbol{\sigma}}\left(\mathbf{x}\right)\right)\odot e^{\boldsymbol{\gamma}}
\end{aligned}
$$

References

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research (JMLR)*, 2(Dec):265–292, 2001.