Supplementary Materials for Coreset Stochastic Variance-Reduced Gradient with Application to Optimal Margin Distribution Machine

Zhi-Hao Tan and **Teng Zhang** and **Wei Wang**
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China
{tanzh, zhangt, wangw}@lamda.nju.edu.cn

In the supplementary material, we will give the detailed proofs of lemma and theorems in main paper. Without loss of generality, we assume that $\|\phi(x)\| = K(x,x)^{1/2} = 1,\ \forall x \in \mathcal{D}$. And we only consider the binary classification, so the label y is either -1 or 1 which implies $|y| = y^2 = 1$.

Bounded Approximation Error

In this section, we will present the detailed proofs of Theorem 1, Lemma 2 and Lemma 3, i.e., the intermediate results when we give the upper bound of approximation error.

Theorem 1. *If $\boldsymbol{w}^*_{odm}$ is the optimal solution of ODM, then there exists a positive constant H such that $\|\boldsymbol{w}^*_{odm}\| \leq H$, where $H = \frac{\sqrt{\lambda(1-\theta)^2+\lambda\mu(1+\theta)^2}}{1-\theta}$. Moreover, for squared hinge loss SVM, the optimal solution $\boldsymbol{w}^*_{h^2}$ satisfies $\boldsymbol{w}^*_{h^2} \leq \lambda$.*

Proof. **1 For ODM**

The optimization problem of ODM can be reformulated as follows:

$$\begin{aligned}\min_{\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon}}\ & \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + \frac{\lambda}{m(1-\theta)^2}\boldsymbol{\xi}^\top\boldsymbol{\xi} + \frac{\lambda\mu}{m(1-\theta)^2}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} \\ \text{s.t. }\ & \boldsymbol{Y}\boldsymbol{x}^\top\boldsymbol{w} \geq (1-\theta)\boldsymbol{e} - \boldsymbol{\xi}, \\ & \boldsymbol{Y}\boldsymbol{x}^\top\boldsymbol{w} \leq (1+\theta)\boldsymbol{e} + \boldsymbol{\epsilon}.\end{aligned} \tag{1}$$

where $\boldsymbol{x}$ is the matrix whose i-th column is $\phi(\boldsymbol{x}_i)$, i.e., $\boldsymbol{x} = [\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_m)]$, $\boldsymbol{Y}$ is a $m \times m$ diagonal matrix with $y_1, \ldots, y_m$ as the diagonal elements and $\boldsymbol{e}$ stands for the all-one vector.

Introduce the Lagrange multipliers $\boldsymbol{\zeta} \geq 0$ and $\boldsymbol{\beta} \geq 0$ for the two constraints respectively, the Largrangian of (1) leads to

$$\begin{aligned}&\mathcal{L}(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon},\boldsymbol{\alpha},\boldsymbol{\beta}) \\ =&\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + \frac{\lambda}{m(1-\theta)^2}\boldsymbol{\xi}^\top\boldsymbol{\xi} + \frac{\lambda\mu}{m(1-\theta)^2}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} \\ &- \boldsymbol{\zeta}^\top\left(\boldsymbol{Y}\boldsymbol{X}^\top\boldsymbol{w} - (1-\theta)\boldsymbol{e} + \boldsymbol{\xi}\right) \\ &+ \boldsymbol{\beta}^\top\left(\boldsymbol{Y}\boldsymbol{X}^\top\boldsymbol{w} - (1+\theta)\boldsymbol{e} - \boldsymbol{\epsilon}\right).\end{aligned} \tag{2}$$

By setting the partial derivative of $\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\epsilon}$ to zero, we have

$$\begin{aligned}&\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} = \boldsymbol{w} - \boldsymbol{X}\boldsymbol{Y}\boldsymbol{\zeta} + \boldsymbol{X}\boldsymbol{Y}\boldsymbol{\beta} = 0 \Longrightarrow \boldsymbol{w} = \boldsymbol{X}\boldsymbol{Y}(\boldsymbol{\zeta} - \boldsymbol{\beta}) \\ &\frac{\partial\mathcal{L}}{\partial\boldsymbol{\xi}} = \frac{2\lambda}{m(1-\theta)^2}\boldsymbol{\xi} - \boldsymbol{\zeta} = 0 \Longrightarrow \boldsymbol{\xi} = \frac{m(1-\theta)^2}{2\lambda}\boldsymbol{\zeta} \\ &\frac{\partial\mathcal{L}}{\partial\boldsymbol{\epsilon}} = \frac{2\lambda\mu}{m(1-\theta)^2}\boldsymbol{\epsilon} - \boldsymbol{\beta} = 0 \Longrightarrow \boldsymbol{\epsilon} = \frac{m(1-\theta)^2}{2\lambda\mu}\boldsymbol{\beta}\end{aligned}$$

Substituting the above to (2), we gain the dual form

$$\begin{aligned}\mathcal{G}(\boldsymbol{\alpha}) = &-\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} - \frac{m(1-\theta)^2}{4\lambda}\boldsymbol{\zeta}^\top\boldsymbol{\zeta} - \frac{m(1-\theta)^2}{4\lambda\mu}\boldsymbol{\beta}^\top\boldsymbol{\beta} \\ &+ (1-\theta)\boldsymbol{\zeta}^\top\boldsymbol{e} - (1+\theta)\boldsymbol{\beta}^\top\boldsymbol{e}\end{aligned} \tag{3}$$

Let us denote $(\boldsymbol{w}^*, \boldsymbol{\xi}^*, \boldsymbol{\epsilon}^*)$ and $(\boldsymbol{\zeta}^*, \boldsymbol{\beta}^*)$ be the primal and dual solutions of (1) and (3), respectively. Since the strong duality holds, we have

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{w}^*\|^2 + \frac{\lambda}{m(1-\theta)^2}\sum_{i=1}^m\left(\xi_i^{*2} + \mu\epsilon_i^{*2}\right) \\ =&-\frac{1}{2}\|\boldsymbol{w}^*\|^2 - \frac{m(1-\theta)^2}{4\lambda}\sum_{i=1}^m\zeta_i^{*2} - \frac{m(1-\theta)^2}{4\lambda\mu}\sum_{i=1}^m\beta_i^{*2} \\ &+ (1-\theta)\sum_{i=1}^m\zeta_i^* - (1+\theta)\sum_{i=1}^m\beta_i^*\end{aligned}$$

Since the loss of ODM is non-negative, we have

$$\begin{aligned}\|\boldsymbol{w}^*\|^2 \leq &-\frac{m(1-\theta)^2}{4\lambda}\sum_{i=1}^m\zeta_i^{*2} - \frac{m(1-\theta)^2}{4\lambda\mu}\sum_{i=1}^m\beta_i^{*2} \\ &+ (1-\theta)\sum_{i=1}^m\zeta_i^* - (1+\theta)\sum_{i=1}^m\beta_i^* \\ =&\sum_{i=1}^m\left((1-\theta)\zeta_i^* - \frac{m(1-\theta)^2}{4\lambda}\zeta_i^{*2}\right) \\ &+ \sum_{i=1}^m\left((1+\theta)\beta_i^* - \frac{m(1-\theta)^2}{4\lambda\mu}\beta_i^{*2}\right) \\ \leq&\sum_{i=1}^m\frac{\lambda}{m} + \sum_{i=1}^m\frac{\lambda\mu(1+\theta)^2}{m(1-\theta)^2}\end{aligned}$$

$$=\frac{\lambda\left(1-\theta\right)^2+\lambda\mu\left(1+\theta\right)^2}{\left(1-\theta\right)^2}=H^2$$

We note that the second inequality uses that the maximum of the quadratic concave function is obtained when $\zeta_i^*=\frac{2\lambda}{m(1-\theta)}$ and $\beta_i^*=\frac{2\lambda\mu(1+\theta)}{m(1-\theta)^2}$. Therefore, for ODM, $\|\boldsymbol{w}_{odm}^*\|\leq H$, where $H=\frac{\sqrt{\lambda(1-\theta)^2+\lambda\mu(1+\theta)^2}}{1-\theta}$.

2 For Square Hinge Loss SVM

The optimization of square hinge loss SVM is as follows:

$$\begin{aligned}&\min_{\boldsymbol{w},\boldsymbol{\xi}}\ \frac{1}{2}\|\boldsymbol{w}\|^2+\frac{\lambda}{m}\sum_{i=1}^m\xi_i^2\\&\text{s.t. } y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)\geq 1-\xi_i,\ i=1,\ldots,m\end{aligned}\tag{4}$$

where $\boldsymbol{\xi}=[\xi_i]_{i=1}^m$. The Lagrange function is of the following form

$$\begin{aligned}\mathcal{L}(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\alpha})=&\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{\lambda}{m}\sum_{i=1}^m\xi_i^2\\&+\sum_{i=1}^m\alpha_i\left(1-\xi_i-y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)\right)\end{aligned}\tag{5}$$

By setting the partial derivative of $\boldsymbol{w}$, $\boldsymbol{\xi}$ to zero, we have

$$\begin{aligned}&\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}}=\boldsymbol{w}-\sum_{i=1}^m\alpha_i y_i\phi(\boldsymbol{x}_i)=0\Longrightarrow\boldsymbol{w}=\sum_{i=1}^m\alpha_i y_i\phi(\boldsymbol{x}_i)\\&\frac{\partial\mathcal{L}}{\partial\xi_i}=\frac{2\lambda}{m}\xi_i-\alpha_i=0\Longrightarrow\xi_i=\frac{m}{2\lambda}\alpha_i\end{aligned}$$

Substituting the above to (5), we gain the dual form

$$\mathcal{W}(\boldsymbol{\alpha})=-\frac{1}{2}\|\boldsymbol{w}\|^2-\frac{m}{4\lambda}\sum_{i=1}^m\alpha_i^2+\sum_{i=1}^m\alpha_i\tag{6}$$

Let us denote $(\boldsymbol{w}^*,\boldsymbol{\xi}^*)$ and $\boldsymbol{\alpha}^*$ be the primal and dual solutions of (4) and (5) respectively. Since the strong duality holds, we have

$$\frac{1}{2}\|\boldsymbol{w}^*\|^2+\frac{\lambda}{m}\sum_{i=1}^m\xi_i^{*2}=-\frac{1}{2}\|\boldsymbol{w}^*\|^2-\frac{m}{4\lambda}\sum_{i=1}^m\alpha_i^{*2}+\sum_{i=1}^m\alpha_i^*$$

Since the loss of square hinge loss SVM is non-negative, we have

$$\begin{aligned}\|\boldsymbol{w}^*\|^2=&-\frac{m}{4\lambda}\sum_{i=1}^m\alpha_i^{*2}+\sum_{i=1}^m\alpha_i^*-\frac{\lambda}{m}\sum_{i=1}^m\xi_i^{*2}\\\leq&\sum_{i=1}^m\left(-\frac{m}{4\lambda}\alpha_i^{*2}+\alpha_i^*\right)\\\leq&\lambda\end{aligned}$$

Therefore, for square hinge loss SVM, we have $\boldsymbol{w}_{h^2}^*\leq\lambda$. In summary, we gain the conclusion. The proof of Theorem 1 is similar to that of Theorem 1 in (Shalev-Shwartz et al. 2011). □

Lemma 2. *For ODM problem, the α_t satisfies $\alpha_t^2\leq A^2$ for all t, where $A=\frac{2(H+1+\theta)}{(1-\theta)^2}$.*

Proof. We first remind the representation of α_t

$$\begin{aligned}\alpha_t=&\frac{2\lambda}{(1-\theta)^2}\{\left(y_t\boldsymbol{w}^\top\phi\left(\boldsymbol{x}_t\right)+\theta-1\right)y_t\mathbb{I}\left(t\in I_1\right)\\&+\mu\left(y_t\boldsymbol{w}^\top\phi\left(\boldsymbol{x}_t\right)-\theta-1\right)y_t\mathbb{I}\left(t\in I_2\right)\}\end{aligned}$$

where $I_1\equiv\left\{i\mid y_t\boldsymbol{w}^\top\phi\left(\boldsymbol{x}_t\right)<1-\theta\right\}$ and $I_2\equiv\left\{i\mid y_t\boldsymbol{w}^\top\phi\left(\boldsymbol{x}_t\right)>1+\theta\right\}$. Then we can obtain

$$\begin{aligned}\|\alpha_t\|\leq&\frac{2\lambda}{\left(1-\theta\right)^2}\left(\left\|\left(y_t\boldsymbol{w}_t^\top\phi\left(\boldsymbol{x}_t\right)+\theta-1\right)\right\|\mathbb{I}(t\in I_2)\right.\\&\left.+\mu\left\|\left(y_t\boldsymbol{w}_t^\top\phi\left(\boldsymbol{x}_t\right)-\theta-1\right)\right\|\mathbb{I}(t\in I_3)\right)\\\leq&\frac{2\lambda}{\left(1-\theta\right)^2}\left(\left(\left\|\boldsymbol{w}_t^\top\phi\left(\boldsymbol{x}_t\right)\right\|+\|\theta-1\|\right)\mathbb{I}(t\in I_2)\right.\\&\left.+\mu\left(\left\|\boldsymbol{w}_t^\top\phi\left(\boldsymbol{x}_t\right)\right\|+\|\theta+1\|\right)\mathbb{I}(t\in I_3)\right)\\\leq&\frac{2\lambda}{\left(1-\theta\right)^2}\left(\left(\|\boldsymbol{w}_t\|\left\|\phi\left(\boldsymbol{x}_t\right)\right\|+1-\theta\right)\mathbb{I}(t\in I_2)\right.\\&\left.+\mu\left(\|\boldsymbol{w}_t\|\left\|\phi\left(\boldsymbol{x}_t\right)\right\|+1+\theta\right)\mathbb{I}(t\in I_3)\right)\\=&\frac{2\lambda}{\left(1-\theta\right)^2}\left(\left(\|\boldsymbol{w}_t\|+1-\theta\right)\mathbb{I}(t\in I_2)\right.\\&\left.+\mu\left(\|\boldsymbol{w}_t\|+1+\theta\right)\mathbb{I}(t\in I_3)\right)\\\leq&\frac{2\lambda}{\left(1-\theta\right)^2}\left(\|\boldsymbol{w}_t\|+1+\theta\right)\\\leq&\frac{2\lambda\left(H+1+\theta\right)}{\left(1-\theta\right)^2}\end{aligned}$$

Note that the first and second inequalities use Minkowski inequality, the third inequality uses Cauchy-Schwarz inequality, and the fourth inequality uses that $\mu\leq 1$ and $0\leq\theta\leq 1$. Therefore, we have $\alpha_t^2\leq A^2$, where $A=\frac{2(H+1+\theta)}{(1-\theta)^2}$. □

Lemma 3. *For hinge loss SVM and square hinge loss SVM, the α_t satisfies $\alpha_t^2\leq A^2=\lambda^2B^2$ for all t, where $B=1$ and $B=2\lambda+2$ respectively.*

Proof. For hinge loss SVM, we have

$$\begin{aligned}l\left(\boldsymbol{w};\boldsymbol{x},y\right)&=\max\left\{0,1-y\boldsymbol{w}^\top\phi\left(\boldsymbol{x}\right)\right\}\\l'\left(\boldsymbol{w};\boldsymbol{x},y\right)&=-\mathbb{I}_{\{y\boldsymbol{w}^\top\phi(\boldsymbol{x})\leq 1\}}y\phi\left(\boldsymbol{x}\right)\end{aligned}$$

where $\mathbb{I}_S$ is the indicator function, which equals 1 if the logical statement S is true and 0 otherwise. Therefore, by taking $B=1$, we have

$$\left\|l'\left(\boldsymbol{w};\boldsymbol{x},y\right)\right\|\leq\left\|\phi\left(\boldsymbol{x}\right)\right\|\leq 1=B$$

For squared hinge loss SVM, we have

$$\begin{aligned}l\left(\boldsymbol{w};\boldsymbol{x},y\right)&=\max\left\{0,1-y\boldsymbol{w}^\top\phi\left(\boldsymbol{x}\right)\right\}^2\\l'\left(\boldsymbol{w};\boldsymbol{x},y\right)&=-\mathbb{I}_{\{y\boldsymbol{w}^\top\phi(\boldsymbol{x})\leq 1\}}2y\phi\left(\boldsymbol{x}\right)\left(1-y\boldsymbol{w}^\top\phi\left(\boldsymbol{x}\right)\right)\end{aligned}$$

Then by taking $B = 2\lambda + 2$ we can obtain

$$\begin{aligned}\|l'(\boldsymbol{w};\boldsymbol{x},y)\| &\le \left\|2y\phi(\boldsymbol{x})\left(1-y\boldsymbol{w}^\top\phi(\boldsymbol{x})\right)\right\| \\ &\le 2\left\|1-y\boldsymbol{w}^\top\phi(\boldsymbol{x})\right\|\|\phi(\boldsymbol{x})\||y| \\ &\le 2\left\|y\boldsymbol{w}^\top\phi(\boldsymbol{x})\right\|+2 \\ &\le 2\|\phi(\boldsymbol{x})\|\|\boldsymbol{w}\||y|+2 \\ &= 2\|\boldsymbol{w}\|+2 \\ &\le 2\lambda+2=B\end{aligned}$$

The last inequality uses Theorem 1.

Therefore, there exists a positive constant B such that $\|l'(\boldsymbol{w};\boldsymbol{x},y)\| \le B$ for hinge loss SVM and square hinge loss SVM. Based on this, we have the following proof.

Note that $\|\phi(\boldsymbol{x})\| = K(x,x) = 1$, by taking $A = \lambda B$, we have

$$\alpha_t^2 = \alpha_t^2 K(\boldsymbol{x}_t,\boldsymbol{x}_t) = \lambda^2\|l'(\boldsymbol{w}_t;\boldsymbol{x}_t,y_t)\|^2 \le \lambda^2 B^2 = A^2$$

Therefore we gain the conclusion $\alpha_t^2 \le A^2$. □

Based on te above results, we obtain the Theorem 4, whose proof is in the main paper.

Theorem 4. *Assume that the p.s.d. and isotropic kernel $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = k(\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)$ is used, where $k(.)$ is a monotonically continuous decreasing function with $k(0) = 1$. And let δ be the diameter of coreset coverage. Then for the approximation error Δ_t as indicated in (10), we have $\|\Delta_t\| \le \frac{3}{2}A\delta_\phi$, where $\delta_\phi = 2\sqrt{2(1-\kappa(\delta^2/4))}$.*

Convergence Analysis

In this section, we will present the proof of intermediate conclusions when analysing the convergence of CSVRG. The proofs of Lemma 6,7 are similar to the proof of Lemma 21, Lemma 22 in (Le et al. 2017).

Lemma 5. *When using CSVRG to train ODM, we have $\mathbb{E}\left[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2\right] \le W^2$ for all t, where $W = 2H$.*

Proof. According to Theorem 1, by taking $W = 2H$, we have the following

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2\right] &\le 2\mathbb{E}\left[\|\boldsymbol{w}_t\|^2\right]+2\mathbb{E}\left[\|\boldsymbol{w}^*\|^2\right] \\ &\le 4H^2 = W^2\end{aligned}$$

□

Lemma 6. *When using CSVRG to train hinge loss SVM or square hinge loss SVM, there exists a positive constant P such that $\mathbb{E}\left[\|\boldsymbol{w}_t\|^2\right] \le P^2$ for all t, where $P = 2A+\frac{3}{2}A\delta_\phi$.*

Proof. We prove by induction that $\mathbb{E}\left[\|\boldsymbol{w}_t\|^2\right] \le P^2$ where $P = 2A+\frac{3}{2}A\delta_\phi$ for all t. Assume that the claim is holding for $t-1$, we have the detailed proof in Figure 1.

The third inequality uses $\sqrt{\mathbb{E}\|\boldsymbol{a}+\boldsymbol{b}\|^2} \le \sqrt{\mathbb{E}\|\boldsymbol{a}\|^2}+\sqrt{\mathbb{E}\|\boldsymbol{b}\|^2}$. The fourth inequality uses Lemma 3. And the fifth inequality uses $\mathbb{E}\|\boldsymbol{\xi}-\mathbb{E}\boldsymbol{\xi}\|^2 = \mathbb{E}\|\boldsymbol{\xi}\|^2-\|\mathbb{E}\boldsymbol{\xi}\|^2 \le \mathbb{E}\|\boldsymbol{\xi}\|^2$ for any random vector $\boldsymbol{\xi}$. Finally, the sixth inequality uses Theorem 4. □

Lemma 7. *Assume that $f(\boldsymbol{w})$ is ν-strongly convex, when using CSVRG to train hinge loss SVM or square hinge loss SVM, we have $\mathbb{E}\left[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2\right] \le W^2$ for all t, where $W = \frac{3A\delta_\phi+\sqrt{9A^2\delta_\phi^2+16(1-\eta\nu)P^2}}{2\nu}$.*

Proof. We first remind the definitions of the relevant variables

$$\begin{aligned}\boldsymbol{g} &= \boldsymbol{w}+\lambda l'(\boldsymbol{w};\boldsymbol{x}_t,y_t) = \boldsymbol{w}+\alpha_t\phi(\boldsymbol{x}_t) \\ \boldsymbol{v}_t &= \nabla\psi_t(\boldsymbol{w}_{t-1})-\nabla\psi_t(\widetilde{\boldsymbol{w}})+\widetilde{\boldsymbol{\mu}} \\ &= \boldsymbol{w}_{t-1}+(\alpha_t-\widetilde{\alpha}_t)\phi(\boldsymbol{x}_t)+\frac{1}{m}\sum_{i=1}^m\widetilde{\alpha}_i\phi(\boldsymbol{x}_i) \\ \Delta_t &= (\alpha_t-\widetilde{\alpha}_t)[\phi(\boldsymbol{c}_t)-\phi(\boldsymbol{x}_t)] \\ &\quad+\frac{1}{m}\sum_{i=1}^m\widetilde{\alpha}_i[\phi(\boldsymbol{c}_i)-\phi(\boldsymbol{x}_i)] \\ \boldsymbol{h}_t &= \boldsymbol{v}_t+\Delta_t\end{aligned}$$

Let $\boldsymbol{d}_t = \boldsymbol{h}_t-\boldsymbol{w}_{t-1}$, according to the proof of lemma 2, we have

$$\begin{aligned}(1-\eta)\sqrt{\mathbb{E}\|\boldsymbol{w}_{t-1}\|^2}+\eta\sqrt{\mathbb{E}\|\boldsymbol{d}_t\|^2} &\le P \\ \Longrightarrow \sqrt{\mathbb{E}\|\boldsymbol{d}_t\|^2} &\le P\end{aligned}$$

Therefore, we can obtain

$$\sqrt{\mathbb{E}\|\boldsymbol{h}_t\|^2} \le \sqrt{\mathbb{E}\|\boldsymbol{w}_{t-1}\|^2}+\sqrt{\mathbb{E}\|\boldsymbol{d}_t\|^2} \le 2P \quad (7)$$

Conditioned on $\boldsymbol{w}_{t-1}$, we have $\mathbb{E}[\boldsymbol{v}_t] = \nabla f(\boldsymbol{w}_{t-1})$. Then we can obtain

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2\right] = \mathbb{E}\left\|\prod_S(\boldsymbol{w}_{t-1}-\eta\boldsymbol{h}_t)-\boldsymbol{w}^*\right\|^2 \\ \le&\mathbb{E}\left[\|\boldsymbol{w}_{t-1}-\eta\boldsymbol{h}_t-\boldsymbol{w}^*\|^2\right] \\ =&\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\|^2+\eta^2\mathbb{E}\left[\|\boldsymbol{h}_t\|^2\right]-2\eta\mathbb{E}[\langle\boldsymbol{w}_{t-1}-\boldsymbol{w}^*,\boldsymbol{v}_t\rangle] \\ &-2\eta\mathbb{E}[\langle\boldsymbol{w}_{t-1}-\boldsymbol{w}^*,\Delta_t\rangle] \\ =&\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\|^2+\eta^2\mathbb{E}\left[\|\boldsymbol{h}_t\|^2\right]-2\eta(\boldsymbol{w}_{t-1}-\boldsymbol{w}^*)^\top\mathbb{E}[\boldsymbol{v}_t] \\ &+2\eta\mathbb{E}\left[\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\|^2\|\Delta_t\|^2\right]^{1/2} \\ \le&\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\|^2+\eta^2\mathbb{E}\left[\|\boldsymbol{h}_t\|^2\right]-\eta\nu\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\|^2 \\ &+2\eta\mathbb{E}\left[\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\|^2\|\Delta_t\|^2\right]^{1/2}\end{aligned}$$

where $S = \mathcal{B}(\boldsymbol{0},\lambda)$. The last inequality uses the ν-strongly convexity of $f(\boldsymbol{w})$ as follows

$$\begin{aligned}&(\boldsymbol{w}_t-\boldsymbol{w}^*)^\top\nabla f(\boldsymbol{w}_t) \\ \ge& f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)+\frac{\lambda}{2}\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2 \\ \ge&\frac{\lambda}{2}\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2\end{aligned}$$

$$
\begin{aligned}
\sqrt{\mathbb{E}\left[\|\boldsymbol{w}_t\|^2\right]} \leq & \sqrt{\mathbb{E}\left[\left\|\prod_S\left(\boldsymbol{w}_{t-1}-\eta \boldsymbol{h}_t\right)\right\|^2\right]} \leq \sqrt{\mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}-\eta \boldsymbol{h}_t\right\|^2\right]} \\
= & \sqrt{\mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}-\eta\left(\boldsymbol{w}_{t-1}+\lambda \ell'\left(\boldsymbol{w}_{t-1} ; x_t, y_t\right)-\widetilde{\boldsymbol{g}}_t+\widetilde{\boldsymbol{\mu}}+\Delta_t\right)\right\|^2\right]} \\
\leq & (1-\eta) \sqrt{\mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}\right\|^2\right]}+\eta \lambda \sqrt{\mathbb{E}\left[\left\|\ell'\left(\boldsymbol{w}_{t-1} ; x_t, y_t\right)\right\|^2\right]}+\eta \sqrt{\mathbb{E}\left[\left\|\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{g}}_t\right\|^2\right]} \\
& +\eta \sqrt{\mathbb{E}\left[\left\|\Delta_t\right\|^2\right]} \\
\leq & (1-\eta) \sqrt{\mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}\right\|^2\right]}+\eta A+\eta \sqrt{\mathbb{E}\left\|\widetilde{\alpha}_t \phi\left(\boldsymbol{x}_t\right)-\frac{1}{m} \sum_{i=1}^m \widetilde{\alpha}_i \phi\left(\boldsymbol{x}_t\right)\right\|^2} \\
& +\eta \sqrt{\mathbb{E}\left[\left\|\Delta_t\right\|^2\right]} \\
\leq & (1-\eta) \sqrt{\mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}\right\|^2\right]}+\eta A+\eta \sqrt{\mathbb{E}\left\|\widetilde{\alpha}_t \phi\left(\boldsymbol{x}_t\right)\right\|^2}+\eta \sqrt{\mathbb{E}\left[\left\|\Delta_t\right\|^2\right]} \\
\leq & (1-\eta) \sqrt{\mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}\right\|^2\right]}+2 \eta A+\frac{3}{2} \eta A \delta_\phi \\
\leq & (1-\eta) P+2 \eta A+\frac{3}{2} \eta A \delta_\phi=P
\end{aligned}
$$

Figure 1: Proof of Lemma 6

The last inequality uses $f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right) \geq 0$. Then by taking expectation again and substituting (4), we have

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\boldsymbol{w}_t-\boldsymbol{w}^*\right\|^2\right] \\
\leq & (1-\eta \nu) \mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\right\|^2\right]+4 \eta^2 P^2 \\
& \quad+2 \eta \mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\right\|^2\left\|\Delta_t\right\|^2\right]^{1 / 2} \\
\leq & (1-\eta \nu) \mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\right\|^2\right]+4 \eta^2 P^2 \\
& \quad+3 \eta A \delta_\phi \mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\right\|^2\right]^{1 / 2}
\end{aligned}
$$

We prove by induction in t. Choosing $W=\frac{3 A \delta_\phi+\sqrt{9 A^2 \delta_\phi^2+16(1-\eta \nu) P^2}}{2 \nu}$, which is the solution of following equation

$$
(1-\eta \nu) W^2+4 \eta^2 P^2+3 \eta A \delta_\phi W=W^2
$$

Therefore, assuming that $\mathbb{E}\left[\left\|\boldsymbol{w}_{t-1}-\boldsymbol{w}^*\right\|^2\right] \leq W^2$, we obtain

$$
\mathbb{E}\left[\left\|\boldsymbol{w}_t-\boldsymbol{w}^*\right\|^2\right] \leq(1-\eta \nu) W^2+4 \eta^2 P^2+3 \eta A \delta_\phi W=W^2
$$

Therefore, we gain the conclusion $\mathbb{E}\left[\left\|\boldsymbol{w}_t-\boldsymbol{w}^*\right\|^2\right] \leq W^2$ for all t. □

According to the Lemma 5, 6, 7, we can obtain the convergence rate as follows, the proof of which is in the main paper.

Theorem 8. *Consider CSVRG in Algorithm 1 with option II and use it to solve SVMs and ODM. Assume that all $\psi_i(\boldsymbol{w})$ are convex and L-smooth, $f(\boldsymbol{w})$ is ν-strongly convex. Let $\boldsymbol{w}^* = argmin_{\boldsymbol{w}} f(\boldsymbol{w})$. Assume that T is sufficiently large so that*

$$
\rho=\frac{1}{\nu \eta(1-4 L \eta) T}+\frac{4 L \eta}{1-4 L \eta}<1
$$

then we have linear convergence in expectation for CSVRG:

$$
\mathbb{E}[f(\widetilde{\boldsymbol{w}}_s)-f(\boldsymbol{w}^*)] \leq \rho^s \mathbb{E}[f(\widetilde{\boldsymbol{w}}_0)-f(\boldsymbol{w}^*)]+\frac{1-\rho^s}{1-\rho} \Omega
$$

where Ω is a constant gap caused by coreset approximation, and $\Omega \to 0$ when the radius of coverage approaches 0.

References

Le, T.; Nguyen, T. D.; Nguyen, V.; and Phung, D. 2017. Approximation vector machines for large-scale online learning. *The Journal of Machine Learning Research* 18(1):3962–4016.

Shalev-Shwartz, S.; Singer, Y.; Srebro, N.; and Cotter, A. 2011. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming* 127(1):3–30.