Decomposition Algorithms for Training Large-scale Semiparametric Support Vector Machines

Sangkyun Lee and Stephen J. Wright

Department of Computer Sciences, University of Wisconsin-Madison, USA

Abstract: We describe a method for solving large-scale semiparametric SVMs for regression problems. Most of the approaches proposed to date for large-scale SVMs cannot accommodate the multiple equality constraints that appear in semiparametric problems. Our approach uses a decomposition framework, with a primal-dual algorithm to find an approximate saddle point for the min-max formulation of each subproblem. We compare our method with algorithms previously proposed for semiparametric SVMs, and show that it scales well as the number of training examples grows.

Semiparametric SVR (SP-SVR)

SVR with ϵ-insensitive loss:

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}||\boldsymbol{w}||_2^2 + \frac{C}{M}\sum_{i=1}^{M} \ell_\epsilon(h;\boldsymbol{x}_i,\boldsymbol{y}_i),$$

$$\ell_\epsilon(h;\boldsymbol{x}_i,\boldsymbol{y}_i) := \max\{|\boldsymbol{y}_i - h(\boldsymbol{x}_i)| - \epsilon, 0\}.$$

- Dataset $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, $i = 1,\dots,M$,
  - $\boldsymbol{x}_i \in \mathbb{R}^N$, $\boldsymbol{y}_i \in \mathbb{R}$.
- Kernel $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) := \langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle$.

Standard (nonparametric) SVR:

$$h(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle + b, \quad b \in \mathbb{R}.$$

Semiparametric SVR:

$$h(\boldsymbol{x}) = \underbrace{\langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle}_{\text{Nonparametric}} + \underbrace{\sum_{j=1}^{K} \beta_j \psi_j(\boldsymbol{x})}_{\text{Parametric}}.$$

Primal Formulation

$$\min_{\boldsymbol{w},\beta,\xi,\xi^*} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{C}{M}\sum_{i=1}^{M}(\xi_i + \xi_i^*) \tag{1a}$$

$$\text{s.t.} \quad \boldsymbol{y}_i - \langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle - \sum_{j=1}^{K}\beta_j\psi_j(\boldsymbol{x}_i) \le \epsilon + \xi_i, \quad i = 1,\dots,M \tag{1b}$$

$$-\left[\boldsymbol{y}_i - \langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle - \sum_{j=1}^{K}\beta_j\psi_j(\boldsymbol{x}_i)\right] \le \epsilon + \xi_i^*, \quad i = 1,\dots,M \tag{1c}$$

$$\xi \ge \boldsymbol{0},\ \xi^* \ge \boldsymbol{0}\ . \tag{1d}$$

Dual Formulation

$$\min_{\boldsymbol{z}} \quad F(\boldsymbol{z}) := \frac{1}{2}\boldsymbol{z}^T\boldsymbol{Q}\boldsymbol{z} + \boldsymbol{p}^T\boldsymbol{z}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}, \qquad (2)$$
$$\boldsymbol{0} \le \boldsymbol{z} \le (C/M)\boldsymbol{1},$$

- $n := 2M$,
- $\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{1} = \{1,\dots,1\}^T \in \mathbb{R}^n$,
- $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ p.s.d.,
- $\boldsymbol{A} \in \mathbb{R}^{K\times n}$, $K \ge 1$.

$\boldsymbol{z} = \begin{bmatrix}\alpha\\ \alpha^*\end{bmatrix} \in \mathbb{R}^{2M}$ for the dual vectors of (1b) and (1c),

$\boldsymbol{p} = [\epsilon - \boldsymbol{y}_1,\dots,\epsilon - \boldsymbol{y}_M, \epsilon + \boldsymbol{y}_1,\dots,\epsilon + \boldsymbol{y}_M]^T \in \mathbb{R}^{2M}$,

$$\boldsymbol{Q}_{ij} = \begin{cases} \boldsymbol{y}_i\boldsymbol{y}_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) & \text{if } 1 \le i,j \le M, \text{ or } M+1 \le i,j \le 2M \\ -\boldsymbol{y}_i\boldsymbol{y}_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) & \text{otherwise} \end{cases},$$

$$\boldsymbol{A} = \begin{bmatrix} \psi_1(\boldsymbol{x}_1) & \cdots & \psi_1(\boldsymbol{x}_M) & -\psi_1(\boldsymbol{x}_1) & \cdots & -\psi_1(\boldsymbol{x}_M) \\ \psi_2(\boldsymbol{x}_1) & \cdots & \psi_2(\boldsymbol{x}_M) & -\psi_2(\boldsymbol{x}_1) & \cdots & -\psi_2(\boldsymbol{x}_M) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \psi_K(\boldsymbol{x}_1) & \cdots & \psi_K(\boldsymbol{x}_M) & -\psi_K(\boldsymbol{x}_1) & \cdots & -\psi_K(\boldsymbol{x}_M) \end{bmatrix} \in \mathbb{R}^{K\times 2M}.$$

Most solvers need $K = 1$

SVM^{light} [3], GPDT [7, 8]
- Working-set selection is for $K = 1$.
- Subproblem solver requires $K = 1$.

SMO [6], LIBSVM [1]
- Working-set selection assumes $K = 1$.
- The analytic solutions of the subproblems are available only when $K = 1$.

Decomposition Algorithm

- In each outer iteration, split variables $\boldsymbol{z}$ into
  - Basic variables $\boldsymbol{z}_\mathcal{B}$, $\mathcal{B} \subset \{1,2,\dots,n\}$.
  - Nonbasic variables $\boldsymbol{z}_\mathcal{N}$, $\mathcal{N} = \{1,2,\dots,n\} \setminus \mathcal{B}$.
  - $\mathcal{B}$ is our working set, of which the size $n_\mathcal{B} \ll n$.
- Fix $\boldsymbol{z}_\mathcal{N}$, change $\boldsymbol{z}_\mathcal{B}$.
- Given $\boldsymbol{z}^k = (\boldsymbol{z}_\mathcal{B}^k, \boldsymbol{z}_\mathcal{N}^k)$, solve the subproblem

$$\min_{\boldsymbol{z}_\mathcal{B}} \quad f(\boldsymbol{z}_\mathcal{B}) := \frac{1}{2}\boldsymbol{z}_\mathcal{B}^T\boldsymbol{Q}_{\mathcal{B}\mathcal{B}}\boldsymbol{z}_\mathcal{B} + (\boldsymbol{Q}_{\mathcal{B}\mathcal{N}}\boldsymbol{z}_\mathcal{N}^k + \boldsymbol{p}_\mathcal{B})^T\boldsymbol{z}_\mathcal{B}$$
$$\text{s.t.} \quad \boldsymbol{A}_\mathcal{B}\boldsymbol{z}_\mathcal{B} = -\boldsymbol{A}_\mathcal{N}\boldsymbol{z}_\mathcal{N}^k + \boldsymbol{b},$$
$$0 \le \boldsymbol{z}_\mathcal{B} \le (C/M)\boldsymbol{1}.$$

- $\boldsymbol{z}^{k+1} \leftarrow (\boldsymbol{z}_\mathcal{B}^{k+1}, \boldsymbol{z}_\mathcal{N}^k)$.

I. Working-set Selection

- $n_\mathcal{B}$: working set size.
- n_c: max. number of "fresh" indices, $\ll n_\mathcal{B}$.

Consider Lagrangian relaxation $\mathcal{L}$ of (2),

$$\mathcal{L}(\boldsymbol{z};\eta) = F(\boldsymbol{z}) + \eta^T\boldsymbol{A}\boldsymbol{z}\ .$$

Given $(\boldsymbol{z}^k, \eta^k)$, find a solution $\boldsymbol{d}$ of

$$\min_{\boldsymbol{d}} \quad \left(\nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}^k;\eta^k)\right)^T\boldsymbol{d}$$
$$\text{s.t.} \quad \begin{aligned} 0 \le \boldsymbol{d}_i \le 1 \quad & \text{if } \boldsymbol{z}_i^{k+1} = 0, \\ -1 \le \boldsymbol{d}_i \le 0 \quad & \text{if } \boldsymbol{z}_i^{k+1} = C/M, \\ -1 \le \boldsymbol{d}_i \le 1 \quad & \text{if } \boldsymbol{z}_i^{k+1} \in (0, C/M), \\ \#\{\boldsymbol{d}_i | \boldsymbol{d}_i \ne 0\} \ \le\ & n_c. \end{aligned} \tag{3}$$

- Solved efficiently in $\mathcal{O}(n\log n)$.
- Convergence proofs in [5, 10].

II. Primal-dual Subproblem Solver (PDSG)

Consider

$$\max_{\eta}\min_{\boldsymbol{z}_\mathcal{B}\in\Omega} \quad \tilde{\mathcal{L}}(\boldsymbol{z}_\mathcal{B},\eta),$$

where

$$\tilde{\mathcal{L}}(\boldsymbol{z}_\mathcal{B},\eta) := f(\boldsymbol{z}_\mathcal{B}) + \eta^T(\boldsymbol{A}_\mathcal{B}\boldsymbol{z}_\mathcal{B} + \boldsymbol{A}_\mathcal{N}\boldsymbol{z}_\mathcal{N}^k)\ ,$$
$$\Omega = \{\boldsymbol{z}_\mathcal{B} \in \mathbb{R}^{n_\mathcal{B}} | \boldsymbol{0} \le \boldsymbol{z}_\mathcal{B} \le (C/M)\boldsymbol{1}\}\ .$$

In each inner iteration, update primal and dual variables by,

$$\begin{cases} \boldsymbol{z}_\mathcal{B}^{\ell+1} \leftarrow \boldsymbol{z}_\mathcal{B}^\ell + s(\boldsymbol{z}_\mathcal{B}^\ell, \eta^\ell) \\ \eta^{\ell+1} \leftarrow \eta^\ell + t(\boldsymbol{z}_\mathcal{B}^{\ell+1}, \eta^\ell)\ , \end{cases}$$

- $s(\cdot,\cdot)$: two-metric GP [2] followed by line-search, on a sub-workingset of size 2.
- $t(\cdot,\cdot)$: $\nabla_\eta\tilde{\mathcal{L}}$, scaled by dual Hessian diagonal [4], on a sub-workingset of size 2.

III. Update

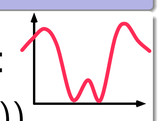
- Update primal-dual iterate pair $(\boldsymbol{z}^{k+1}, \eta^{k+1})$.
  - $\boldsymbol{z}^{k+1} \leftarrow (\boldsymbol{z}_\mathcal{B}^{k+1}, \boldsymbol{z}_\mathcal{N}^k)$.
  - η^{k+1} is provided by the subproblem solver.
- Incremental full gradient $\nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z};\eta)$ update,

$$\nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}^{k+1},\eta^{k+1}) = \nabla F(\boldsymbol{z}^k) + \begin{bmatrix}\boldsymbol{Q}_{\mathcal{B}\mathcal{B}}\\ \boldsymbol{Q}_{\mathcal{N}\mathcal{B}}\end{bmatrix}(\boldsymbol{z}_\mathcal{B}^{k+1} - \boldsymbol{z}_\mathcal{B}^k) + (\eta^{k+1})^T\boldsymbol{A}.$$

Experiment: Toy Problem

Modified Mexican hat function [4, 9]:

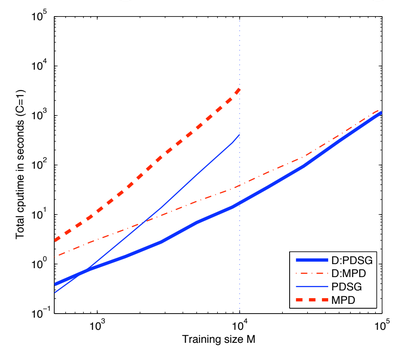
$$\omega(x) = \sin(x) + \operatorname{sinc}(2\pi(x-5)).$$



- $y_i = \omega(x_i) + \zeta_i$: $x_i \sim$ Uniform$[0,10]$, $\zeta_i \sim \mathcal{N}(0, 0.2^2)$.
- $\psi_1(x) = \sin(x), \psi_2(x) = \operatorname{sinc}(2\pi(x-5))$.
- Gaussian kernel $\kappa(x,y) = \exp(-0.25||x-y||^2)$.
- $\epsilon = 0.05$, $n_\mathcal{B} = 500$, $n_c = 100$.

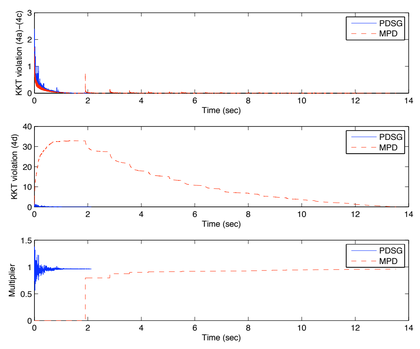
Compare to the current best solver MPD [4].
- No decomposition.
- Primal-dual method (method of multipliers).
- Gradient projection (primal), scaled gradient ascent (dual) on (sub-)workingset of size 1.

Scaling w.r.t. Training Size



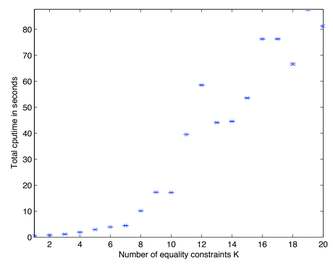
- PDSG vs. MPD (stand-alone).
- D:PDSG vs. D:MPD (in decomposition).
- D:MPD catches up D:PDSG when $M\uparrow$: the full gradient update step becomes dominant as $M\uparrow$.

Convergence ($M = 1000$)



- PDSG vs. MPD (stand-alone).
- (T) max. dual feasibility violation.
- (M) max. primal equality constraints violation.
- (B) the first parametric component coeff.

Scaling of D:PDSG w.r.t. K ($M = 1000$)



- Time complexity of D:PDSG is $\mathcal{O}(uKn_\mathcal{B})$, u is the no. of outer iterations.
- Solver time seems to increase linearly with K for $K \ge 6$.

$$\psi_j(x) = \begin{cases}\cos(j\pi x), j = 0,2,4,\dots \\ \sin(j\pi x),\ j = 1,3,5,\dots\end{cases}$$

Experiment: Real Dataset

Air pollution effect on respiratory illness in Milan, Italy, '80–'89 [11].
- 3652 daily records of env. conditions and the no. of respiratory deaths. Features are rescaled to $[0,1]$.
- Semiparametric model to predict the no. of deaths,

$$h_{\text{sp}}(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle + \beta_1(\boldsymbol{x}_{\text{temp}}) + \beta_2(\boldsymbol{x}_{\text{SO}_2}) + \beta_3(\boldsymbol{x}_{\text{temp}})^2 + \beta_4(\boldsymbol{x}_{\text{SO}_2})^2 + \beta_5,$$

- The fitted model is,

$$h_{\text{sp}}(\boldsymbol{x}) = \langle \boldsymbol{w}^*, \phi(\boldsymbol{x})\rangle + 0.22(\boldsymbol{x}_{\text{temp}} - 0.47)^2 + 0.26(\boldsymbol{x}_{\text{SO}_2}) - 0.07(\boldsymbol{x}_{\text{SO}_2})^2 + 0.12.$$

  - Deaths are lower in the middle of the temp. range.
  - Linear increase of death rate with SO_2 level.

Summary

- Introduced an efficient algorithm for solving a generalized dual formulation of SVMs, which has multiple equality constraints.
- Easily extends to other SVMs, e.g. ν-SVMs, semiparametric extensions of SVMs, etc.
- Future work: theoretical analysis, fine-tuning working-set selection, stochastic approximation alternative method.

http://www.cs.wisc.edu/~sklee

sklee@cs.wisc.edu

Decomposition Algorithms for Training Large-scale Semiparametric Support Vector Machines

Sangkyun Lee and Stephen J. Wright

Department of Computer Sciences, University of Wisconsin-Madison, USA

[1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, April 2009. version 2.89, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[2] E. M. Gafni and D. P. Bertsekas. Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22:936–964, 1984.

[3] Thorsten Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[4] Wolf Kienzle and Bernhard Schölkopf. Training support vector machines with multiple equality constraints. In *Machine Learning: ECML 2005*, volume 16, October 2005.

[5] C. J. Lin. Linear convergence of a decomposition method for support vector machines. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2001.

[6] John C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 12, pages 185–208. MIT Press, Cambridge, MA, 1999.

[7] T. Serafini, G. Zanghirati, and L. Zanni. Gradient projection methods for large quadratic programs and applications in training support vector machines. *Optimization Methods and Software*, 20(2–3):353–378, 2004.

[8] T. Serafini and L. Zanni. On the working set selection in gradient projection-based decomposition techniques for support vector machines. *Optimization Methods and Software*, 20:583–596, 2005.

[9] Alex J. Smola, Thilo T. Frieß, and Bernhard Schölkopf. Semiparametric support vector and linear programming machines. In *Advances in Neural Information Processing Systems 11*, pages 585–591, Cambridge, MA, USA, 1999. MIT Press.

[10] P. Tseng and S. Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. Published online in Computational Optimization and Applications, October 2008.

[11] M A Vigotti, G Rossi, L Bisanti, A Zanobetti, and J Schwartz. Short term effects of urban air pollution on respiratory health in Milan, Italy, 1980-89. *Journal of Epidemiology Community Health*, 50:s71–75, 1996.

http://www.cs.wisc.edu/~sklee sklee@cs.wisc.edu