Advanced Topics in Machine Learning

1. Learning SVMs / Bundle Methods

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim, Germany

Outline

6. Cutting Plane Algorithm

7. Digression: Bundle Methods

8. Bundle Methods for SVMs

Outline

Structural SVM

$$\min f(\hat{\beta}, \hat{\xi}) := \frac{1}{2}\hat{\beta}^T\hat{\beta} + \gamma n\hat{\xi} \qquad \text{[STRUCT.SVM]}$$

$$\text{w.r.t.} \frac{1}{n}\sum_{i=1}^{n} c_i y_i \hat{\beta}^T x_i \geq \frac{||c||_1}{n} - \hat{\xi}, \quad c \in \mathcal{C}$$

$$\hat{\xi} \geq 0$$

for given $\gamma > 0$ and $\mathcal{C} \subseteq \{0,1\}^n$.

Equivalence to LSVM

Lemma

The (original) linear SVM problem [LSVM] and the structured SVM problem [STRUCT.SVM] for $\mathcal{C} := \{0,1\}^n$ are equivalent.

Proof.

"⇒": Let $(\hat{\beta}, \hat{\xi})$ be a feasible point of [LSVM]. Then $(\hat{\beta}, \tilde{\xi})$ with $\tilde{\xi} := \frac{1}{n}\sum_{i=1}^n \xi_i$ is feasible for [STRUCT.SVM]: for any $c \in \mathcal{C}$:

$$\frac{1}{n}\sum_{i=1}^n c_i y_i \hat{\beta}^T x_i \geq \frac{1}{n}\sum_i c_i(1-\xi_i) \geq \frac{1}{n}\sum_i c_i - \frac{1}{n}\sum_i \xi_i = \frac{||c||_1}{n} - \tilde{\xi}$$

and $f_{\text{LSVM}}(\hat{\beta}, \hat{\xi}) = \frac{1}{2}\hat{\beta}^T\hat{\beta} + \gamma\sum_{i=1}^n \hat{\xi}_i = \frac{1}{2}\hat{\beta}^T\hat{\beta} + \gamma n\tilde{\xi} = f_{\text{STRUCT.SVM}}(\hat{\beta}, \tilde{\xi})$

Equivalence to LSVM (2/2)

"$\Leftarrow$": Let $(\hat{\beta}, \tilde{\xi})$ be a feasible point of [STRUCT.SVM]. Then $(\hat{\beta}, \hat{\xi})$ with

$$\tilde{\xi}_i := [1 - y_i \hat{\beta}^T x_i]_+$$

is feasible for [LSVM].
Now let

$$c := (\delta_{1 - y_i \hat{\beta}^T x_i > 0})_{i=1,\dots,n}$$

Then

$$\sum_{i=1}^{n} \tilde{\xi}_i = \sum_{i=1}^{n} c_i (1 - y_i \hat{\beta}^T x_i) \leq n\tilde{\xi}$$

and thus $f_{\text{LSVM}}(\hat{\beta}, \hat{\xi}) = \frac{1}{2}\hat{\beta}^T \hat{\beta} + \gamma \sum_{i=1}^{n} \hat{\xi}_i \leq \frac{1}{2}\hat{\beta}^T \hat{\beta} + \gamma n \tilde{\xi} = f_{\text{STRUCT.SVM}}(\hat{\beta}, \tilde{\xi}$

Dual Formulation

Lemma

The dual formulation of [STRUCT.SVM] is

$$\max \ \bar{f}(\hat{\alpha}) := -\sum_{c,d\in\mathcal{C}} \hat{\alpha}_c\hat{\alpha}_d q_c^T q_d + \sum_{c\in\mathcal{C}} \frac{||c||_1}{n}\hat{\alpha}_c$$

$$w.r.t. \sum_{c\in\mathcal{C}} \hat{\alpha}_c \leq \gamma$$

$$\hat{\alpha}_c \geq 0, \quad c \in \mathcal{C}$$

with

$$q_c := \frac{1}{n}\sum_{i=1}^{n} c_i y_i x_i$$

Basic Ideas

Basic Ideas:

- start with $\mathcal{C} = \emptyset$.
- In each iteration,
  add the constraint for the set of examples with errors.
- Do not solve the primal structured problem,
  but the dual structured problem
  (only $|\mathcal{C}|$ variables).
- Store q_c.

Initialization

If we start with $\mathcal{C} := \emptyset$ and optimize the primal [STRUCT.SVM], we get

$$\hat{\beta} = 0$$
$$\hat{\xi} = 0$$
$$c = \mathbb{e}$$

Cutting Plane Algorithm (Joachims 2006)

learn-linear-svm-cutting-plane(training predictors x, training targets y,
complexity γ, accuracy ϵ) :

$\mathcal{C} := \{\mathbb{e}\}$

$q_{\mathbb{e}} := \frac{1}{n}\sum_{i=1}^{n} y_i x_i$

do

$\quad \hat{\alpha} := \operatorname{argmax}\left\{ -\frac{1}{2}\sum_{c,d\in\mathcal{C}} \alpha_c \alpha_d q_c^T q_d + \sum_{c\in\mathcal{C}} \frac{||c||_1}{n}\alpha_c \;\middle|\; \sum_{c\in\mathcal{C}} \alpha_c \leq \gamma, \alpha_c \geq 0 \quad \forall c \in \mathcal{C} \right\}$

$\quad \hat{\beta} := \sum_{c\in\mathcal{C}} \hat{\alpha}_c q_c$

$\quad \hat{\xi} := \max_{c\in\mathcal{C}} \frac{||c||_1}{n} - \hat{\beta}^T q_c$

$\quad c := (\delta_{y_i \hat{\beta}^T x_i < 1})_{i=1,\ldots,n}$

$\quad q_c := \frac{1}{n}\sum_{i=1}^{n} c_i y_i x_i$

$\quad \mathcal{C} := \mathcal{C} \cup \{c\}$

while $\frac{||c||_1}{n} - \hat{\beta}^T q_c > \hat{\xi} + \epsilon$

return $\hat{\beta}$

Outline

Derivatives as Linear Approximation (Fréchet Derivative)

Definition (Fréchet derivative)

Let $f : U \to Y$ be a function on an open subset $U \subseteq X$ of a Banach space X into a Banach space Y. f is called Fréchet differentiable at $x \in U$ if there is a bounded linear operator $A_x : X \to Y$ with

$$\lim_{h \to 0} \frac{||f(x+h) - f(x) - A_x(h)||_Y}{||h||_X} = 0$$

Then $Df(x) := A_x$ is called its Fréchet derivative at x.

Banach space: complete normed vector space (i.e., contains the limit of every Cauchy sequence).
$\mathbb{R}^n$ with Euclidean norm is a Banach space.

Bounded linear operator A**:** exists $M \in \mathbb{R}_0^+$ with $||Ax||_Y \leq M||x||_X$ for every x.
For finite dimensional spaces all linear operators are bounded.
Example unbounded linear operator: X the vector space of all bounded sequences in $\mathbb{R}$ with norm $||x|| := \sup\{x_i \mid i \in \mathbb{N}\}$.
Then $A : X \to X$ with $A(x) := (i\, x_i)_{i \in \mathbb{N}}$ is linear, but not bounded.

Derivatives as Linear Approximation (Fréchet Derivative)

The Fréchet derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be described by the Jacobian matrix:

$$Df(x) = A_x = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

If f is Fréchet differentiable at x, it is continuous at x.

Directional Derivatives & Gâteaux Derivative

Definition (Directional derivative)

Let $f : U \to Y$ be a function on an open subset $U \subseteq X$ of a Banach space X into a Banach space Y. f is called differentiable at $x \in U$ in direction $d \in X$ if

$$df(x; d) := \lim_{t \searrow 0} \frac{f(x + td) - f(x)}{t}$$

exists. Then $df(x; d)$ is called its derivative at x in direction d.

Note: Directional and Gâteaux derivatives are defined for more general spaces, so called locally convex topological vector spaces.

Directional Derivatives & Gâteaux Derivative

Definition (Gâteaux derivative)

If the derivative of f at x in direction d exists for every d and is linear in d, f is called Gâteaux differentiable at x.
If X is a Hilbert space and thus

$$df(x; d) = \langle a, d \rangle, \quad \text{for an } a \in X$$

then $\nabla_x f := a$ is called Gâteaux derivative.

If f is Fréchet differentiable at x, then it also is Gâteaux differentiable at x and both derivatives coincide.
The reverse is not true.

Hilbert space: real or complex vector space with inner product, that is complete w.r.t. metric induced by inner product.
Every Hilbert space is a Banach space.

Directional Derivatives & Gâteaux Derivative

Derivatives in all directions may exist, but fail to depend linearly on the direction.

Example:

$$f(x,y) := \begin{cases} \frac{x^3}{x^2+y^2}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{else} \end{cases}$$

has derivative at $(0,0)$ in every direction d

$$df(x;d) := \lim_{t\to 0} \frac{f(x+td)-f(x)}{t} = \lim_{t\to 0} \frac{\frac{t^3 d_1^3}{t^2 d_1^2 + t^2 d_2^2}}{t} = \frac{d_1^3}{d_1^2+d_2^2}$$

but df is not linear in d, i.e., f not Gâteaux differentiable.

Gâteaux vs. Fréchet Derivative

Also non-continuous functions may be Gâteaux differentiable.

Example:

$$f(x,y) := \begin{cases} \frac{x^3 y}{x^6+y^2}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{else} \end{cases}$$

is non-continuous at $(0,0)$,
but its derivative at $(0,0)$ in direction d is

$$df(x;d) := \lim_{t\to 0} \frac{f(x+td)-f(x)}{t} = \lim_{t\to 0} \frac{\frac{t^3 d_1^3 t d_2}{t^6 d_1^6 + t^2 d_2^2}}{t}$$

$$= \lim_{t\to 0} \frac{t d_1^3 d_2}{t^4 d_1^6 + d_2^2} = 0$$

thus linear in d and Gâteaux differentiable.

Gâteaux vs. Fréchet Derivative

Even a continuous function may be Gâteaux differentiable, but not Fréchet differentiable.

Example:

$$f(x,y) := \begin{cases} \frac{x^2 y}{x^4+y^2}\sqrt{x^2+y^2}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{else} \end{cases}$$

is continuous at $(0,0)$ and Gâteaux differentiable with derivative 0, but not Fréchet differentiable as

$$\lim_{h \to 0} \frac{||f(x+h) - f(x) - A_x(h)||_Y}{||h||_X} = \lim_{h \to 0} \frac{||f(h)||_Y}{||h||_X}$$

along $h = (t, t^2)$

$$= \lim_{t \to 0} \frac{t^2 t^2}{t^4 + t^4} \sqrt{t^2 + t^4} / \sqrt{t^2 + t^4} = \frac{1}{2} \neq 0$$

Subgradients

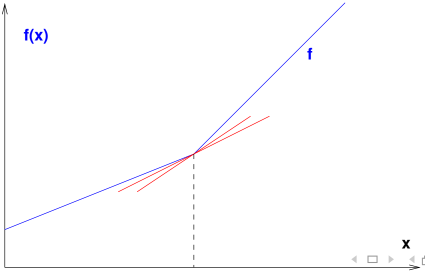
Definition (Subgradients)

Let $f : U \to \mathbb{R}$ be a function on an open subset $U \subseteq X$ of a Hilbert space X.

A vector $\phi \in X$ is called a subgradient of f at x if

$$\langle \phi, \tilde{x} - x \rangle \leq f(\tilde{x}) - f(x), \quad \forall \tilde{x} \in U$$

The set of all subgradients of f at x is called its subdifferential $\partial_x f$ at x.



Subgradients vs. Directional Derivatives

- If f is convex, then

$$\phi \in \partial_x f \quad \Longleftrightarrow \quad \langle \phi, \cdot \rangle \leq df(x; \cdot)$$

- If f is convex, then

$$df(x; d) = \max_{\phi \in \partial_x f} \langle d, \phi \rangle$$

- If f is convex, then

$$f \text{ is Gâteaux differentiable at } x \quad \Longleftrightarrow \quad |\partial_x f| = 1$$

and then $\partial_x f = \{\nabla_x f\}$.

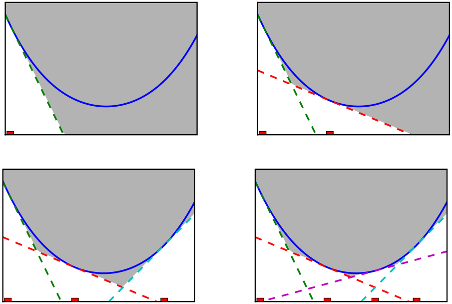
Cutting Plane Method

The cutting plane method approximates a function by a sequence of its subgradients $\phi_t \in \partial_{x_t} f$ at different iterates x_t:

$$f(x) \geq f(x_t) + \langle \phi_t, x - x_t \rangle, \quad \forall x \in U$$

and thus

$$f(x) \geq f^{(t)}(x) := \max_{t'=1,\ldots,t} f(x_{t'}) + \langle \phi_{t'}, x - x_{t'} \rangle, \quad \forall x \in U$$



[Teo et al. 2009]

Generic Cutting Plane Algorithm

minimize-cutting-plane(function f) :
choose (randomly) $x_0 \in \text{dom} f$
compute $\phi_0 \in \partial_{x_0} f$
$a_0 := f(x_0) - \langle \phi_0, x_0 \rangle$
$t := 0$
while $||\phi_t|| > 0$ **do**
    $x_{t+1} := \text{argmin}_x f^t := \text{argmin}_x \max_{t'=1,\ldots,t} a_{t'} + \langle \phi_{t'}, x \rangle$
    compute $\phi_{t+1} \in \partial_{x_{t+1}} f$
    $a_{t+1} := f(x_{t+1}) - \langle \phi_{t+1}, x_{t+1} \rangle$
    $t := t + 1$
od
return x_t

Generic Cutting Plane Algorithm

Variant with line search:

minimize-cutting-plane-line-search(function f) :
choose (randomly) $x_0 \in \text{dom} f$
compute $\phi_0 \in \partial_{x_0} f$
$a_0 := f(x_0) - \langle \phi_0, x_0 \rangle$
$t := 0$
while $||\phi_t|| > 0$ do
    $\tilde{x}_{t+1} := \text{argmin}_x\, f^t := \text{argmin}_x \max_{t'=1,\ldots,t} a_{t'} + \langle \phi_{t'}, x \rangle$
    $\eta := \text{argmin}_\eta\, f(x_t + \eta(\tilde{x}_{t+1} - x_t))$
    $x_{t+1} := x_t + \eta(\tilde{x}_{t+1} - x_t)$
    compute $\phi_{t+1} \in \partial_{x_{t+1}} f$
    $a_{t+1} := f(x_{t+1}) - \langle \phi_{t+1}, x_{t+1} \rangle$
    $t := t + 1$
od
return x_t

Bundle Methods

Control step size by a proximity control function:

- proximal bundle methods (Kiwiel 1990):

$$\tilde{x}_{t+1} := \arg\min_x f^t(x) + \frac{\zeta_t}{2}||x - x_t||^2$$

- trust region bundle methods (Schramm and Zowe 1992):

$$\tilde{x}_{t+1} := \arg\min\{f^t(x) \mid x \text{ with } \frac{1}{2}||x - x_t||^2 \leq \kappa_t\}$$

- level set bundle methods (Lemaréchal et al. 1995):

$$\tilde{x}_{t+1} := \arg\min\{\frac{1}{2}||x - x_t||^2 \mid x \text{ with } f^t(x) \leq \tau_t\}$$

Bundle Methods / Subproblems in the Dual

The subproblems (with ζ_t a constant)

$$x_{t+1} := \arg\min_x \left(\max_{t'=1,\dots,t} a_{t'} + \langle \phi_{t'}, x \rangle \right) + \frac{\zeta_t}{2} ||x - x_t||^2$$

can be solved in the dual:

$$\alpha := \arg\max_\alpha -\frac{1}{2\zeta_t} \alpha^T \Phi \Phi^T \alpha + b^T \alpha$$

$$\text{w.r.t. } \mathbb{e}^T \alpha = 1$$

$$\alpha \geq 0$$

where

$$\Phi := \begin{pmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_t^T \end{pmatrix}, \quad b := a + \Phi x_t$$

Then

$$x = x_t - \frac{1}{\zeta_t} \Phi^T \alpha$$

Bundle Methods / Subproblems in the Dual

Proof.

$$\arg\min_x \tilde{f}(x) := \xi + \frac{\zeta_t}{2}||x - x_t||^2$$
$$\text{w.r.t. } \xi \geq a_{t'} + \langle \phi_{t'}, x \rangle, \quad t' = 1, \ldots, t$$

Lagrange function
$$F_{\tilde{f}}(x, \xi, \alpha) = \xi + \frac{\zeta_t}{2}||x - x_t||^2 + \alpha^T(a + \Phi x - \xi \mathbb{e})$$
$$= \xi(1 - \alpha^T \mathbb{e}) + \frac{\zeta_t}{2}||x - x_t||^2 + \alpha^T \Phi x + \alpha^T a$$

$$\frac{\partial F_{\tilde{f}}}{\partial x} = \zeta^t(x - x_t) + \alpha^T \Phi \stackrel{!}{=} 0 \qquad \rightsquigarrow x = x_t - \frac{1}{\zeta_t}\Phi^T \alpha \ (I)$$
$$\frac{\partial F_{\tilde{f}}}{\partial \xi} = (1 - \alpha^T \mathbb{e}) \stackrel{!}{=} 0 \qquad \rightsquigarrow \mathbb{e}^T \alpha = 1 \ (II)$$

Bundle Methods / Subproblems in the Dual

Proof (ctd.).

$$\bar{f}(\alpha) := \inf_{x,\xi} F_{\tilde{f}}(x,\xi,\alpha) = \frac{\zeta_t}{2\zeta_t^2}\alpha^T\Phi\Phi^T\alpha + \alpha^T\Phi(x_t - \frac{1}{\zeta_t}\Phi^T\alpha) + \alpha^T a$$
$$= -\frac{1}{2\zeta_t}\alpha^T\Phi\Phi^T\alpha + \alpha^T(\Phi x_t + a)$$

Outline

A Slightly Different Problem Formulation

The classical SVM literature formulation (with C instead of γ):

$$\begin{aligned}\text{minimize } f(\beta, \beta_0, \xi) &:= \frac{1}{2}||\beta||^2 + \gamma\langle \mathbb{e}, \xi\rangle \\ \text{w.r.t. } y \odot (\beta_0 \mathbb{e} + X\beta) &\geq \mathbb{e} - \xi \\ \xi &\geq 0\end{aligned}$$

The risk & regularization formulation:

$$\begin{aligned}\text{minimize } f(\beta, \beta_0, \xi) &:= \frac{1}{n}\langle \mathbb{e}, \xi\rangle + \frac{1}{2}\lambda||\beta||^2 \\ \text{w.r.t. } y \odot (\beta_0 \mathbb{e} + X\beta) &\geq \mathbb{e} - \xi \\ \xi &\geq 0\end{aligned}$$

obviously are equivalent for

$$\lambda = \frac{1}{n\gamma}$$

Subgradient for Risk on Hinge Loss

Bundle methods can be applied to non-differential risks, such as risk on Hinge loss:

$$R(\hat{\beta}, \hat{\beta}_0; x, y) := \frac{1}{n} \sum_{i=1}^{n} [1 - y_i(\hat{\beta}^T x_i + \hat{\beta}_0)]_+$$

with subgradient

$$g(\hat{\beta}, \hat{\beta}_0; x, y) := \begin{pmatrix} -\frac{1}{n} \sum\limits_{\substack{i=1: \\ y_i(\hat{\beta}^T x_i + \hat{\beta}_0) < 1}}^{n} y_i x_i \\ -\frac{1}{n} \sum\limits_{\substack{i=1 \\ y_i(\hat{\beta}^T x_i + \hat{\beta}_0) < 1}}^{n} y_i \end{pmatrix}$$

Bundle Methods and L2 Regularization

Teo et al. 2009 see structural similarities between the proximity control of proximal bundle methods and the L2 regularization term. Differences:

- Always penalize relative to $x_t = 0$
- Use $\zeta_t := \lambda$ as weight.
- x is called β, t' is called i, $\phi_{t'}$ is called $-q_i$.

$$\tilde{\beta}_{t+1} := \arg\min_x f^t(\beta) + \frac{\lambda}{2}||\beta||^2$$

or in the dual

$$\alpha := \arg\max_\alpha -\frac{1}{2\lambda}\alpha^T Q Q^T \alpha + a^T \alpha$$

$$\text{w.r.t. } \mathbb{e}^T \alpha = 1$$
$$\alpha \geq 0$$

Cutting Plane Algorithm and L2 Regularization

Alternatively, one could extend the Cutting Plane Algorithm slightly to handle functions of type

$$f(x) := f_1(x) + f_2(x)$$

where f_1 is non-differentiable, but f_2 is. Then approximate f by

$$f^{(t)}(x) := \max_{t'=1,\ldots,t} f_1(x_t) + g_t^T(x - x_t) + f_2(x)$$

with $g_t \in \partial_{x_t} f_1$.

Loss functions and their derivatives

Table 5: Scalar loss functions and their derivatives, depending on $f := \langle w, x \rangle$, and y.

	Loss $\bar{l}(f, y)$	Derivative $\bar{l}'(f, y)$
Hinge (Bennett and Mangasarian, 1992)	$\max(0, 1 - yf)$	0 if $yf \geq 1$ and $-y$ otherwise
Squared Hinge (Keerthi and DeCoste, 2005)	$\frac{1}{2}\max(0, 1 - yf)^2$	0 if $yf \geq 1$ and $f - y$ otherwise
Exponential (Cowell et al., 1999)	$\exp(-yf)$	$-y\exp(-yf)$
Logistic (Collins et al., 2000)	$\log(1 + \exp(-yf))$	$-y/(1 + \exp(-yf))$
Novelty (Schölkopf et al., 2001)	$\max(0, \rho - f)$	0 if $f \geq \rho$ and -1 otherwise
Least mean squares (Williams, 1998)	$\frac{1}{2}(f - y)^2$	$f - y$
Least absolute deviation	$\lvert f - y \rvert$	$\mathrm{sgn}(f - y)$
Quantile regression (Koenker, 2005)	$\max(\tau(f - y), (1 - \tau)(y - f))$	τ if $f > y$ and $\tau - 1$ otherwise
ϵ-insensitive (Vapnik et al., 1997)	$\max(0, \lvert f - y \rvert - \epsilon)$	0 if $\lvert f - y \rvert \leq \epsilon$, else $\mathrm{sgn}(f - y)$
Huber's robust loss (Müller et al., 1997)	$\frac{1}{2}(f - y)^2$ if $\lvert f - y \rvert \leq 1$, else $\lvert f - y \rvert - \frac{1}{2}$	$f - y$ if $\lvert f - y \rvert \leq 1$, else $\mathrm{sgn}(f - y)$
Poisson regression (Cressie, 1993)	$\exp(f) - yf$	$\exp(f) - y$

Table 6: Vectorial loss functions and their derivatives, depending on the vector $f := Wx$ and on y.

	Loss	Derivative
Soft-Margin Multiclass (Taskar et al., 2004) (Crammer and Singer, 2003)	$\max_{y'}(f_{y'} - f_y + \Delta(y, y'))$	$e_{y^*} - e_y$ where y^* is the argmax of the loss
Scaled Soft-Margin Multiclass (Tsochantaridis et al., 2005)	$\max_{y'} \Gamma(y, y')(f_{y'} - f_y + \Delta(y, y'))$	$\Gamma(y, y')(e_{y^*} - e_y)$ where y^* is the argmax of the loss
Softmax Multiclass (Cowell et al., 1999)	$\log \sum_{y'} \exp(f_{y'}) - f_y$	$\left[\sum_{y'} e_{y'} \exp(f'_y)\right] / \sum_{y'} \exp(f'_y) - e_y$
Multivariate Regression	$\frac{1}{2}(f - y)^\top M(f - y)$ where $M \succeq 0$	$M(f - y)$

[Teo et al. 2009]

References

Joachims, Thorsten (2006): *Training linear SVMs in linear time*. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '06. ACM ID: 1150429. Philadelphia, PA, USA: ACM, 217–226.

Teo, C. H et al. (2009): *Bundle methods for regularized risk minimization*. In: *Journal of Machine Learning Research* 1, 1–55.