

Tikhonov Regularization, Value Regularization, and Fenchel Duality

Ryan M. Rifkin

Honda Research Institute USA, Inc.
Human Intention Understanding Group

2007

HONDA
The Power of Dreams



- Data points $X = \{X_1, \dots, X_n\}$. $X_i \in \mathbb{R}^d$. X can also refer to an n by d matrix (the *rows* are the points).
- Labels $Y = \{Y_1, \dots, Y_n\}$. $Y_i \in \mathbb{R}$. Y can also refer to an n -vector. Note that the labels are *capital* Y .

Notation: Loss Functions

- We will use V or v to denote loss functions.
- Usually, V will be a loss function operating on an entire data set, and v will be the pointwise loss.
- We will be cavalier about hiding the data in the definition of the loss function. So you may see, for the square loss:

$$V(Y, f(X)) = \frac{1}{2} \|Y - f(X)\|^2$$

$$V(f) = \frac{1}{2} \|Y - f(X)\|^2$$

$$v_i(Y_i, f(X_i)) = \frac{1}{2} \|Y_i - f(X_i)\|^2$$

$$v_i(f) = \frac{1}{2} \|Y_i - f(X_i)\|^2.$$

Tikhonov Regularization Revisited

- Tikhonov regularization:

$$\min_{f \in \mathcal{H}} V(f(X), Y) + \frac{\lambda}{2} \|f\|_K^2.$$

- By the representer theorem, the solution has the form

$$f(\cdot) = \sum_i c_i k(X_i, \cdot)$$

- Therefore, what we are really looking for is

$$\min_{c \in \mathbb{R}^n} V(Kc, Y) + \frac{\lambda}{2} c^t Kc.$$

Value Regularization

- Assume K^{-1} exists.
- In Tikhonov Regularization, we are finding a function. Define y to be the values of that function on the training set:

$$y = f(X) = Kc \quad (\implies c = K^{-1}y)$$

- Instead of optimizing c , we optimize y :

$$\begin{aligned} & \min_{c \in \mathbb{R}^n} V(Kc, Y) + \lambda c^t Kc \\ = & \min_{y \in \mathbb{R}^n} V(KK^{-1}y, Y) + \lambda y^t K^{-1}KK^{-1}y \\ = & \min_{y \in \mathbb{R}^n} V(y, Y) + \frac{\lambda}{2} y^t K^{-1}y. \end{aligned}$$

- This is a simple change with far reaching consequences.

Value Regularization vs. Coefficient Regularization

- Two different approaches:

$$\min_{c \in \mathbb{R}^n} V(Kc, Y) + \frac{\lambda}{2} c^t K c$$
$$\min_{y \in \mathbb{R}^n} V(y, Y) + \frac{\lambda}{2} y^t K^{-1} y$$

- Note that in the value formulation, the kernel function appears only in the regularizer, not in the loss.
- It is very natural to write the value form as

$$\min_{y \in \mathbb{R}^n} \sum_i v_i(y_i, Y_i) + \lambda y^t K^{-1} y.$$

We can do the same thing for the coefficient formulation, but each v_i will depend on all the c 's.

Fenchel Duality, Main Theorem (Reminder)

Theorem

Given convex functions f and g , under minor technical conditions,

$$\inf_{y,z} \{f(y) + g(y) + f^*(z) + g^*(-z)\} = 0,$$

at least one minimizer exists, and all minimizers y, z satisfy the complementarity equations:

$$\begin{aligned} f(y) - y^t z + f^*(z) &= 0 \\ g(y) + y^t z + g^*(-z) &= 0. \end{aligned}$$

Value Regularization, Fenchel Duality Optimality

- Define the regularization function

$$R(y) = \frac{\lambda}{2} y^t K^{-1} y.$$

- In value regularization, we are looking for

$$\min_{y \in \mathbb{R}^n} R(y) + V(y).$$

- Using the Fenchel duality theorem, we are looking for a y and z that satisfy

$$\begin{aligned} R(y) - y^t z + R^*(y) &= 0 \\ V(y) + y^t z + V^*(-z) &= 0. \end{aligned}$$

- We have a *separation of concerns*: the regularizer and the loss contribute separate optimality conditions.

Regularization Optimality Condition

- We are looking for y and z satisfying

$$R(y) - y^t z + R^*(z) = 0.$$

- For Tikhonov regularization,

$$\begin{aligned}R(y) &= \lambda y^t K^{-1} y. \\R^*(z) &= \lambda^{-1} z^t K z.\end{aligned}$$

- The optimality condition for the regularizer is:

$$\begin{aligned}\frac{1}{2} \lambda y^t K^{-1} y - y^t z + \frac{1}{2} \lambda^{-1} z^t K z &= 0 \\ \frac{1}{2} (y - \lambda^{-1} K z)^t (\lambda K^{-1} y - z) &= 0 \\ y = \lambda^{-1} K z &\iff z = \lambda K^{-1} y.\end{aligned}$$

Regularization Optimality Condition

- For Tikhonov regularization, the optimal y and z satisfy

$$y = \lambda^{-1} Kz,$$

independent of the loss function.

- Modified regularizers will lead to modified optimality conditions, again independent of the loss. Key future example: unregularized bias terms.
- The z 's are closely related to the expansion coefficients via $c = \lambda y$.

Loss Optimality Conditions

- For a pointwise loss function

$$V(y) = \sum_i v_i(y_i),$$

the conjugate of the sum is the sum of the conjugates:

$$\begin{aligned} V^*(z) &= \sup_y \left\{ y^t z - \sum_i v_i(y_i) \right\} \\ &= \sum_i \sup_{y_i} \{ y_i z_i - v_i(y_i) \} \\ &= \sum_i v_i^*(z_i). \end{aligned}$$

- Therefore, for each data point, we get a constraint

$$v_i(y_i) + y_i z_i + v_i^*(-z_i).$$

The exact form of the constraint is dictated by the loss.