Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 10 - Optimization Foundations Applied to Regression Formulations

Building on questions on Least Squares Linear Regression

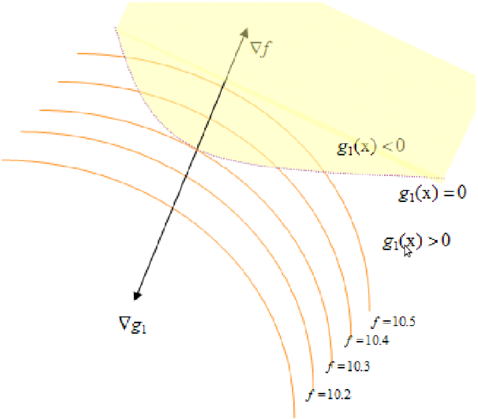
1. Is there a probabilistic interpretation?
   - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
   - Bayesian and Maximum Aposteriori Estimates, Regularization, Support Vector Regression
3. How to minimize the resultant and more complex error functions?
   - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

Constrained convex problems

Q. *How to solve such constrained problems?*

A. Canonical example:

$$\text{Minimize } f(\mathbf{w}) \ s.t. \ g_1(\mathbf{w}) \leq 0 \tag{1}$$



Constrained Convex Problems

- If $\mathbf{w}^*$ is on the boundary of g_1, *i.e.*, if $g_1(\mathbf{w}^*) = 0$,

$$\nabla f(\mathbf{w}^*) = -\lambda \nabla g_1(\mathbf{w}^*) \textit{ for some } \lambda \geq 0$$

- **Intuition:** If the above didn't hold, then we would have $\nabla f(\mathbf{w}^*) = \lambda_1 \nabla g_1(\mathbf{w}^*) + \lambda_2 \nabla_\perp g_1(\mathbf{w}^*)$, where, by moving in direction[1] $\pm\nabla_\perp g_1(\mathbf{w}^*)$ (or $-\nabla g_1(\mathbf{w}^*)$), we remain on boundary $g_1(\mathbf{w}^*) = 0$, (or within $g_1(\mathbf{w}^*) \leq 0$) while decreasing the value of f, which is not possible at the point of optimality.
- Thus, at the point of optimality[2], for some $\lambda \geq 0$,

$$\textit{Either } g_1(\mathbf{w}^*) < 0 \quad \& \quad \nabla f(\mathbf{w}^*) = 0 \tag{2}$$

$$\textit{Or } g_1(\mathbf{w}^*) = 0 \quad \& \quad \nabla f(\mathbf{w}^*) = -\lambda \nabla g_1(\mathbf{w}^*) \tag{3}$$

[1] $\nabla_\perp g_1(\mathbf{w}^*)$ is the direction orthogonal to $\nabla g_1(\mathbf{w}^*)$

[2] `Section 4.4, pg-72: cs725/notes/BasicsOfConvexOptimization.pdf`

Explaining the Figure

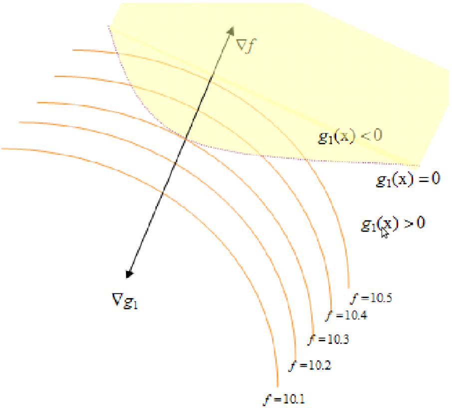


Figure 2: Two conditions under which a minimum can occur: a) When the minimum is on the constraint function boundary, in which case the gradients are in opposite directions; b) When point of minimum is inside the constraint space (shown in yellow shade), in which case $\nabla f(\mathbf{w}^*) = \mathbf{0}$.

More Explanation and Lagrange Function

- The first condition occurs when minima lies on the boundary of function g. In this case, gradient vectors corresponding to the functions f and g, at $\mathbf{w}^*$, point in opposite directions barring multiplication by a real constant.
- Second condition represents the case that point of minimum lies inside the constraint space. This space is shown shaded in Figure 1. Clearly, for this case, $\nabla f(\mathbf{w}) = \mathbf{0}$.
- An Alternative Representation: $\nabla L(\mathbf{w}, \lambda) = 0$ for some $\lambda \geq 0$ where

$$L(\mathbf{w}, \lambda) = f(\mathbf{w}) + \lambda \mathbf{g}(\mathbf{w}); \lambda \in \mathbb{R}$$

is called the lagrange function which has objective function augmented by weighted sum of constraint functions

Duality and KKT conditions

For a convex objective and constraint function, the minima, $\mathbf{w}^*$, can satisfy one of the following two conditions:

1. $g(\mathbf{w}^*) = \mathbf{0}$ and $\nabla f(\mathbf{w}^*) = -\lambda \nabla \mathbf{g}(\mathbf{w}^*)$
2. $g(\mathbf{w}^*) < \mathbf{0}$ and $\nabla f(\mathbf{w}^*) = \mathbf{0}$

Duality and KKT conditions

- Here, we wish to penalize higher magnitude coefficients, hence, we wish $g(\mathbf{w})$ to be negative while minimizing the lagrangian. In order to maintain such direction, we must have $\lambda \geq 0$. Also, for solution $\mathbf{w}$ to be feasible, $\nabla g(\mathbf{w}) \leq \mathbf{0}$.
- Due to complementary slackness condition, we further have $\lambda g(\mathbf{w}) = \mathbf{0}$, which roughly suggests that the lagrange multiplier is zero unless constraint is active at the minimum point. As $\mathbf{w}$ minimizes the lagrangian $L(\mathbf{w}, \lambda)$, gradient must vanish at this point and hence we have $\nabla f(\mathbf{w}) + \lambda \nabla \mathbf{g}(\mathbf{w}) = \mathbf{0}$

KKT Conditions, Duality, SVR Dual

KKT conditions for the Constrained (Convex) Problem

- The general optimization problem we consider with (convex) inequality and (linear) equality constraints is:

$$\min_{\mathbf{w}} f(\mathbf{w})$$

KKT conditions for the Constrained (Convex) Problem

- The general optimization problem we consider with (convex) inequality and (linear) equality constraints is:

$$\min_{\mathbf{w}} f(\mathbf{w})$$

$$\text{subject to } g_i(\mathbf{w}) \leq 0; 1 \leq i \leq m$$

$$h_j(\mathbf{w}) = 0; 1 \leq j \leq p$$

KKT conditions for the Constrained (Convex) Problem

- Here, $\mathbf{w} \in \mathbb{R}^n$ and the domain is the intersection of all functions. Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

KKT conditions for the Constrained (Convex) Problem

- Here, $\mathbf{w} \in \mathbb{R}^n$ and the domain is the intersection of all functions. Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- KKT **necessary** conditions for all differentiable functions (i.e. f, g_i, h_j) with optimality points $\hat{\mathbf{w}}$ and $(\hat{\lambda}, \hat{\mu})$ are:
  - $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
  - $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
  - $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$

KKT conditions for the Constrained (Convex) Problem

- Here, $\mathbf{w} \in \mathbb{R}^n$ and the domain is the intersection of all functions. Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- KKT **necessary** conditions for all differentiable functions (i.e. f, g_i, h_j) with optimality points $\hat{\mathbf{w}}$ and $(\hat{\lambda}, \hat{\mu})$ are:
  - $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
  - $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
  - $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$
- When f and $g_i, \forall i \in [1, m]$ are convex and $h_j, \forall j \in [1, p]$ are affine, KKT conditions are also **sufficient** for optimality at $\hat{\mathbf{w}}$ and $(\hat{\lambda}, \hat{\mu})$

Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

$$L^*(\lambda, \mu) = \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu)$$

- The Dual Optimization Problem is to maximize Lagrange dual function $L^*(\lambda, \mu)$ over (λ, μ)

Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

$$L^*(\lambda, \mu) = \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu)$$

- The Dual Optimization Problem is to maximize Lagrange dual function $L^*(\lambda, \mu)$ over (λ, μ)

$$\underset{\lambda,\mu}{\operatorname{argmax}}\ L^*(\lambda, \mu) = \underset{\lambda,\mu}{\operatorname{argmax}} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu)$$

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

$$\max_{\lambda,\mu} L^*(\lambda, \mu) = \max_{\lambda,\mu} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu) \le L(\mathbf{w}, \lambda, \mu)$$

- **Duality Gap:** The gap between primal and dual solutions. In the KKT conditions, $\hat{\mathbf{w}}$ correspond to primal optimal and $(\hat{\lambda}, \hat{\mu})$ to dual optimal points $\Rightarrow$ Duality gap is $f(\hat{\mathbf{w}}) - L^*(\hat{\lambda}, \hat{\mu})$
- Duality gap characterizes suboptimality of the solution and can be approximated by $f(\mathbf{w}) - L^*(\lambda, \mu)$ for any feasible $\mathbf{w}$ and corresponding λ and μ

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

$$\max_{\lambda,\mu} L^*(\lambda, \mu) = \max_{\lambda,\mu} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu) \leq L(\mathbf{w}, \lambda, \mu)$$

- **Duality Gap:** The gap between primal and dual solutions. In the KKT conditions, $\hat{\mathbf{w}}$ correspond to primal optimal and $(\hat{\lambda}, \hat{\mu})$ to dual optimal points $\Rightarrow$ Duality gap is $f(\hat{\mathbf{w}}) - L^*(\hat{\lambda}, \hat{\mu})$
- Duality gap characterizes suboptimality of the solution and can be approximated by $f(\mathbf{w}) - L^*(\lambda, \mu)$ for any feasible $\mathbf{w}$ and corresponding λ and μ
- When functions f and $g_i, \forall i \in [1, m]$ are convex and $h_j, \forall j \in [1, p]$ are affine, Karush-Kuhn-Tucker (KKT) conditions are both necessary and sufficient for points to be both primal and dual optimal with zero duality gap.

Support Vector Regression and its Dual

Instructor: Prof. Ganesh Ramakrishnan

KKT and Dual for SVR

- $$\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$$
  s.t. $\forall i$,
  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
  $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$,
  $\xi_i, \xi_i^* \geq 0$
- Let's consider the lagrange multipliers α_i, α_i^*, μ_i and μ_i^* corresponding to the above-mentioned constraints.
- The Lagrange Function is

KKT and Dual for SVR

- $$\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$$
  s.t. $\forall i$,
  $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i,$
  $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*,$
  $\xi_i, \xi_i^* \geq 0$
- Let's consider the lagrange multipliers α_i, α_i^*, μ_i and μ_i^* corresponding to the above-mentioned constraints.
- The Lagrange Function is $L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) =$
  $$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
  $$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i (\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

$$\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0 \textit{ i.e., } \mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$

- Differentiating the Lagrangian w.r.t. ξ_i,

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i (\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $$\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0 \text{ } i.e., \text{ } \mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i)$$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
$$\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0 \textit{ i.e., } \mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$
- Differentiating the Lagrangian w.r.t. ξ_i,
$C - \alpha_i - \mu_i = 0$ *i.e.,* $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
$\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^{m}\alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^{m}\alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^{m}\mu_i\xi_i - \sum_{i=1}^{m}\mu_i^*\xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $$\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0 \textit{ i.e., } \mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.,* $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i(\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:

KKT conditions for SVR

$$L(\mathbf{w},\alpha,\alpha^*,\mu,\mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i+\xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

  $$\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0 \text{ i.e., } \mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i(\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ AND $\mu_i\xi_i = 0$ AND
  $\alpha_i^*(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$ AND $\mu_i^*\xi_i^* = 0$

Conclusions from the KKT conditions:

$$\alpha_i \in (0, C) \Rightarrow ?$$

$$\alpha_i^* \in (0, C) \Rightarrow ?$$

KKT conditions

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $w - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$
  i.e. $\mathbf{w} = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$
  i.e. $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i^m (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$
  $\mu_i \xi_i = 0$
  $\alpha_i^* (b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$
  $\mu_i^* \xi_i^* = 0$

Conclusions from the KKT conditions:

$$\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$$

and

$$\alpha_i^*(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$$

$\Rightarrow$?

Conclusions from the KKT conditions:

$$\alpha_i \in (0, C) \Rightarrow ?$$

$$(C - \alpha_i)\xi_i = 0 \Rightarrow ?$$

$$\alpha_i^* \in (0, C) \Rightarrow ?$$

$$(C - \alpha_i^*)\xi_i^* = 0 \Rightarrow ?$$

For Support Vector Regression, since the original objective and the constraints are convex, any $(\mathbf{w}, b, \alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*)$ that satisfy the necessary KKT conditions gives optimality (conditions are also sufficient)

Some observations

- $\alpha_i, \alpha_i^* \geq 0$, $\mu_i, \mu_i^* \geq 0$, $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
  Thus, $\alpha_i, \mu_i, \alpha_i^*, \mu_i^* \in [0, C]$, $\forall i$
- If $0 < \alpha_i < C$, then $0 < \mu_i < C$
  (as $\alpha_i + \mu_i = C$)
- $\mu_i \xi_i = 0$ and $\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ are complementary slackness conditions
  So $0 < \alpha_i < C \Rightarrow \xi_i = 0$ and $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b = \epsilon + \xi_i = \epsilon$
  - All such points lie on the boundary of the ϵ band
  - Using any point $\mathbf{x}_j$ (that is with $\alpha_j \in (0, C)$) on margin, we can recover b as:
    $b = y_j - \mathbf{w}^\top \phi(\mathbf{x}_j) - \epsilon$

KKT Conditions, Duality, SVR Dual

KKT conditions for SVR

$$L(\mathbf{w},\alpha,\alpha^*,\mu,\mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i+\xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $$\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0 \textit{ i.e., } \mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.,* $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i(\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ AND $\mu_i\xi_i = 0$ AND
  $\alpha_i^*(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$ AND $\mu_i^*\xi_i^* = 0$

For Support Vector Regression, since the original objective and the constraints are convex, any $(\mathbf{w}, b, \alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*)$ that satisfy the necessary KKT conditions gives optimality (conditions are also sufficient)

Some observations

- $\alpha_i, \alpha_i^* \geq 0$, $\mu_i, \mu_i^* \geq 0$, $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
  Thus, $\alpha_i, \mu_i, \alpha_i^*, \mu_i^* \in [0, C], \forall i$
- If $0 < \alpha_i < C$, then $0 < \mu_i < C$
  (as $\alpha_i + \mu_i = C$)
- $\mu_i \xi_i = 0$ and $\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ are complementary slackness conditions
  So $0 < \alpha_i < C \Rightarrow \xi_i = 0$ and $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b = \epsilon + \xi_i = \epsilon$
  - All such points lie on the boundary of the ϵ band
  - Using any point $\mathbf{x}_j$ (that is with $\alpha_j \in (0, C)$) on margin, we can recover b as:
    $b = y_j - \mathbf{w}^\top \phi(\mathbf{x}_j) - \epsilon$

Support Vector Regression

Dual Objective

Weak Duality

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, we have:
  $\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$
  s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
  $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
  $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$
- The above is true for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$
- Thus,

Weak Duality

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, we have:
  $\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$
  s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
  $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
  $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$
- The above is true for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$
- Thus,

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \geq \max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

  s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
  $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
  $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$

Dual objective

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- Assume: In case of SVR, we have a strictly convex objective and linear constraints $\Rightarrow$ KKT conditions are necessary and sufficient and strong duality holds:

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) = \max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
$w^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
$\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$

- This value is precisely obtained at the $(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$ that satisfies the necessary (and sufficient) KKT optimality conditions
- Given strong duality, we can equivalently solve

$$\max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

- $L(\alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m(\xi_i + \xi_i^*) +$
  $\sum_{i=1}^{m}\left(\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \alpha_i^*(\mathbf{w}^\top\phi(\mathbf{x}_i) + b - y_i - \epsilon - \xi_i^*)\right) -$
  $\sum_{i=1}^{m}(\mu_i\xi_i + \mu_i^*\xi_i^*)$
- We obtain $\mathbf{w}$, b, ξ_i, ξ_i^* in terms of α, α^*, μ and μ^* by using the KKT conditions derived earlier as $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$ and $\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
- Thus, we get:

- $L(\alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m(\xi_i + \xi_i^*) + \sum_{i=1}^m \left(\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \alpha_i^*(\mathbf{w}^\top\phi(\mathbf{x}_i) + b - y_i - \epsilon - \xi_i^*)\right) - \sum_{i=1}^m(\mu_i\xi_i + \mu_i^*\xi_i^*)$
- We obtain $\mathbf{w}$, b, ξ_i, ξ_i^* in terms of α, α^*, μ and μ^* by using the KKT conditions derived earlier as $\mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$ and $\sum_{i=1}^m(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
- Thus, we get:
$L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
$= \frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j) + \sum_i\left(\xi_i(C - \alpha_i - \mu_i) + \xi_i^*(C - \alpha_i^* - \mu_i^*)\right) - b\sum_i(\alpha_i - \alpha_i^*) - \epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*) - \sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $L(\alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) +$
$\sum_{i=1}^m \left(\alpha_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \alpha_i^* (\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i - \epsilon - \xi_i^*) \right) -$
$\sum_{i=1}^m (\mu_i \xi_i + \mu_i^* \xi_i^*)$
- We obtain $\mathbf{w}$, b, ξ_i, ξ_i^* in terms of α, α^*, μ and μ^* by using the KKT conditions derived earlier as $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i)$ and $\sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
- Thus, we get:
$L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
$= \frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j) +$
$\sum_i \left(\xi_i (C - \alpha_i - \mu_i) + \xi_i^* (C - \alpha_i^* - \mu_i^*) \right) - b \sum_i (\alpha_i - \alpha_i^*) - \epsilon \sum_i (\alpha_i + \alpha_i^*) +$
$\sum_i y_i (\alpha_i - \alpha_i^*) - \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j)$
$= -\frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j) - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*)$

Kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + y_j - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon$ $\mathbf{x}_j$ is any point with $\alpha_j \in (0, C)$. Recall similarity with

Kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + y_j - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon$ $\mathbf{x}_j$ is any point with $\alpha_j \in (0, C)$. Recall similarity with kernelized expression for Ridge Regression
- The dual optimization problem to compute the α's for SVR is:

Kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + y_j - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon$ $\mathbf{x}_j$ is any point with $\alpha_j \in (0, C)$. Recall similarity with kernelized expression for Ridge Regression
- The dual optimization problem to compute the α's for SVR is:

$$max_{\alpha_i, \alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$$
$$-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$$

  s.t.
  - $\sum_i (\alpha_i - \alpha_i^*) = 0$
  - $\alpha_i, \alpha_i^* \in [0, C]$
- **We notice that the only way these three expressions involve ϕ is through $\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$, for some i, j**

Recap from Quiz 1: Kernelizing Ridge Regression

- Given $w = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y$ and using the identity $(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T(BPB^T + R)^{-1}$
  - $\Rightarrow w = \Phi^T(\Phi\Phi^T + \lambda I)^{-1} y = \sum_{i=1}^{m} \alpha_i \phi(x_i)$ where $\alpha_i = \left((\Phi\Phi^T + \lambda I)^{-1} y\right)_i$
  - $\Rightarrow$ the final decision function $f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w} = \sum_{i=1}^{m} \alpha_i \phi^T(\mathbf{x})\phi(\mathbf{x}_i)$
- Again, **We notice that the only way the decision function $f(\mathbf{x})$ involves ϕ is through $\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$, for some i, j**

The Kernel function

- We call $\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$ a **kernel function**: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$
- The Kernel Trick: For some important choices of ϕ, compute $K(\mathbf{x}_i, \mathbf{x}_j)$ directly and more efficiently than having to explicitly compute/enumerate $\phi^($\mathbf{x}_i$) and $\phi(\mathbf{x}_j)$
- The expression for decision function becomes $f(x) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$
- Computation of α_i is specific to the objective function being minimized: Closed form exists for Ridge regression but NOT for SVR

Back to the Kernelized version of SVR

- The kernelized dual problem:

$$max_{\alpha_i,\alpha_i^*} -\frac{1}{2}\sum_i\sum_j(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x}_i,\mathbf{x}_j)$$

$$-\epsilon\sum_i(\alpha_i+\alpha_i^*)+\sum_i y_i(\alpha_i-\alpha_i^*)$$

  s.t.
  - $\sum_i(\alpha_i-\alpha_i^*)=0$
  - $\alpha_i,\alpha_i^*\in[0,C]$
- The kernelized decision function:
  $f(\mathbf{x})=\sum_i(\alpha_i-\alpha_i^*)K(\mathbf{x}_i,\mathbf{x})+b$
- Using any point x_j with $\alpha_j\in(0,C)$:
  $b=y_j-\sum_i(\alpha_i-\alpha_i^*)K(\mathbf{x}_i,\mathbf{x}_j)$
- Computing $K(\mathbf{x}_1,\mathbf{x}_2)$ often does not even require computing $\phi(\mathbf{x}_1)$ or $\phi(\mathbf{x}_2)$ explicitly

Basis function expansion and the Kernel trick

- We started off with the functional form[3]

$$f(\mathbf{x}) = \sum_{j=1}^{p} w_j \phi_j(\mathbf{x})$$

Each ϕ_j is called a *basis function* and this representation is called *basis function expansion*[4]

- And we landed up with an equivalent

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

for Ridge regression and Support Vector Regression

- Aside: For $p \in [0, \infty)$, with what K, kind of regularizers, loss functions, *etc.*, will these dual representations hold?[5]

[3] The additional b term can be either absorbed in ϕ or kept separate as discussed on several occasions.

[4] Section 2.8.3 of Tibshi

[5] Section 5.8.1 of Tibshi

An Example Kernel

- Let $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1^\top \mathbf{x}_2)^2$
- What $\phi(\mathbf{x})$ will give $\phi^\top(\mathbf{x}_1)\phi(\mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1^\top \mathbf{x}_2)^2$
- Is such a ϕ guaranteed to exist?
- Is there a unique ϕ for given K?

An Example Kernel

- We can prove that such a ϕ exists
- For example, for a 2-dimensional $\mathbf{x}_i$:

$$\phi(\mathbf{x}_i) = \begin{bmatrix} 1 \\ x_{i1}\sqrt{2} \\ x_{i2}\sqrt{2} \\ x_{i1}x_{i2}\sqrt{2} \\ x_{i1}^2 \\ x_{i2}^2 \end{bmatrix}$$

- $\phi(\mathbf{x}_i)$ exists in a 5-dimensional space
- But, to compute $K(\mathbf{x}_1, \mathbf{x}_2)$, all we need is $x_1^\top x_2$ without having to enumerate $\phi(\mathbf{x}_i)$

More on the Kernel Trick

- **Kernels** operate in a *high-dimensional*, *implicit* feature space without necessarily computing the coordinates of the data in that space, but rather by simply computing the Kernel function
- This approach is called the "*kernel trick*" and will subsequently talk about *valid kernels*
- This operation is often computationally cheaper than the explicit computation of the coordinates
- Claim: If $\mathcal{K}_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ are entries of an $n \times n$ **Gram Matrix** $\mathcal{K}$ then
  - $\mathcal{K}$ must be positive semi-definite
  - Proof: $\mathbf{b}^T \mathcal{K} \mathbf{b} = \sum_{i,j} b_i \mathcal{K}_{ij} b_j = \sum_{i,j} b_i b_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$

$$= \langle \sum_i b_i \phi(\mathbf{x}_i), \sum_j b_j \phi(\mathbf{x}_j) \rangle = || \sum_i b_i \phi(\mathbf{x}_i) ||_2^2 \geq 0$$

Existence of basis expansion ϕ for symmetric K?

- *Positive-definite kernel:* For any dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and for any m, the Gram matrix $\mathcal{K}$ must be positive definite

$$\mathcal{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \ldots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \ldots & K(\mathbf{x}_i, \mathbf{x}_j) & \ldots \\ K(\mathbf{x}_m, \mathbf{x}_1) & \ldots & K(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

so that $\mathcal{K} = U\Sigma U^T = (U\Sigma^{\frac{1}{2}})(U\Sigma^{\frac{1}{2}})^T = RR^T$ where rows of U are linearly independent and Σ is a positive diagonal matrix

[6] Eigen-decomposition wrt linear operators. See
`https://en.wikipedia.org/wiki/Mercer%27s_theorem`

[7] That is, if every Cauchy sequence is convergent.

Existence of basis expansion ϕ for symmetric K?

- *Positive-definite kernel:* For any dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and for any m, the Gram matrix $\mathcal{K}$ must be positive definite

$$\mathcal{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \ldots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \ldots & K(\mathbf{x}_i, \mathbf{x}_j) & \ldots \\ K(\mathbf{x}_m, \mathbf{x}_1) & \ldots & K(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

  so that $\mathcal{K} = U\Sigma U^T = (U\Sigma^{\frac{1}{2}})(U\Sigma^{\frac{1}{2}})^T = RR^T$ where rows of U are linearly independent and Σ is a positive diagonal matrix
- *Mercer kernel:* Extending to eigenfunction decomposition[6]:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{\infty} \alpha_j \phi_j(\mathbf{x}_1)\phi_j(\mathbf{x}_2) \text{ where } \alpha_j \geq 0 \text{ and } \sum_{j=1}^{\infty} \alpha_j^2 < \infty$$

- *Mercer kernel* and *Positive-definite kernel* turn out to be equivalent if the input space $\{x\}$ is *compact*[7]

---

[6] Eigen-decomposition wrt linear operators. See `https://en.wikipedia.org/wiki/Mercer%27s_theorem`

[7] That is, if every Cauchy sequence is convergent.

Mercer's theorem

- **Mercer kernel:** $K(\mathbf{x}_1, \mathbf{x}_2)$ is a Mercer kernel if $\int\int K(\mathbf{x}_1, \mathbf{x}_2) g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$ for all square integrable functions $g(\mathbf{x})$ ($g(\mathbf{x})$ is square integrable *iff* $\int (g(\mathbf{x}))^2\, dx$ is finite)
- **Mercer's theorem:**
  An implication of the theorem:
  for any *Mercer kernel* $K(\mathbf{x}_1, \mathbf{x}_2)$, $\exists\, \phi(\mathbf{x}) : \mathbb{R}^n \mapsto H$,
  s.t. $K(\mathbf{x}_1, \mathbf{x}_2) = \phi^\top(\mathbf{x}_1)\phi(\mathbf{x}_2)$
  - where H is a *Hilbert space*[8], the infinite dimensional version of the Eucledian space.
  - Eucledian space: $(\Re^n, < .,. >)$ where $< .,. >$ is the standard dot product in $\Re^n$
  - Advanced: Formally, *Hibert Space* is an inner product space with associated norms, where every Cauchy sequence is convergent

[8] Do you know Hilbert? No? Then what are you doing in his space? :)

Prove that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

- We want to prove that
  $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$,
  for all square integrable functions $g(\mathbf{x})$
- Here, $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors s.t $\mathbf{x}_1, \mathbf{x}_2 \in \Re^t$
- Thus, $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2$

$$= \int_{x_{11}} .. \int_{x_{1t}} \int_{x_{21}} .. \int_{x_{2t}} \left[\sum_{n_1..n_t} \frac{d!}{n_1!..n_t!} \prod_{j=1}^{t} (x_{1j} x_{2j})^{n_j} \right] g(x_1) g(x_2)\, dx_{11}..dx_{1t} dx_{21}..dx_{2t}$$

$$\text{s.t. } \sum_{i=1}^{t} n_i = d$$

(taking a leap)

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \prod_{j=1}^{t} (x_{1j} x_{2j})^{n_j} \, g(x_1) g(x_2) \, dx_1 dx_2$$

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (x_{11}^{n_1} x_{12}^{n_2} \ldots x_{1t}^{n_t}) g(x_1) \, (x_{21}^{n_1} x_{22}^{n_2} \ldots x_{2t}^{n_t}) g(x_2) \, dx_1 dx_2$$

Prove that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \prod_{j=1}^{t} (x_{1j} x_{2j})^{n_j} \, g(x_1) g(x_2) \, dx_1 dx_2$$

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (x_{11}^{n_1} x_{12}^{n_2} \ldots x_{1t}^{n_t}) g(x_1) \, (x_{21}^{n_1} x_{22}^{n_2} \ldots x_{2t}^{n_t}) g(x_2) \, dx_1 dx_2$$

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \left(\int_{\mathbf{x}_1} (x_{11}^{n_1} \ldots x_{1t}^{n_t}) g(x_1) \, dx_1 \right) \left(\int_{\mathbf{x}_2} (x_{21}^{n_1} \ldots x_{2t}^{n_t}) g(x_2) \, dx_2 \right)$$

(integral of decomposable product as product of integrals)

$$\text{s.t.} \sum_{i}^{t} n_i = d$$

Prove that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

- Realize that both the integrals are basically the same, with different variable names
- Thus, the equation becomes:

$$\sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \left(\int_{\mathbf{x}_1} (x_{11}^{n_1} \ldots x_{1t}^{n_t}) g(x_1)\, dx_1 \right)^2 \geq 0$$

(the square is non-negative for reals)

- Thus, we have shown that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is a Mercer kernel.

What about $\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ s.t. $\alpha_d \geq 0$?

- $K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$
- Is $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d \right) g(\mathbf{x}_1) g(\mathbf{x}_2) \, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$?
- We have

$$\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d \right) g(x_1) g(x_2) \, dx_1 dx_2 =$$

What about $\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ s.t. $\alpha_d \geq 0$?

- $K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$
- Is $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d \right) g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$?
- We have

$$\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d \right) g(x_1) g(x_2)\, dx_1 dx_2 =$$

$$\sum_{d=1}^{r} \alpha_d \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2$$

What about $\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ s.t. $\alpha_d \geq 0$?

- We have already proved that $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2) \, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$
- Also, $\alpha_d \geq 0$, $\forall d$
- Thus,

$$\sum_{d=1}^{r} \alpha_d \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (\mathbf{x}1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2) \, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$$

- By which, $K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ is a Mercer kernel.
- Examples of Mercer Kernels: Linear Kernel, Polynomial Kernel, Radial Basis Function Kernel

Kernels in SVR

- Recall:
  $max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i\sum_j(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x}_i,\mathbf{x}_j) - \epsilon\sum_i(\alpha_i+\alpha_i^*) + \sum_i y_i(\alpha_i-\alpha_i^*)$
  and the decision function:
  $f(x) = \sum_i(\alpha_i-\alpha_i^*)K(\mathbf{x}_i,\mathbf{x}) + b$
  are all in terms of the kernel $K(\mathbf{x}_i,\mathbf{x}_j)$ only
- *One can now employ any mercer kernel in SVR or Ridge Regression to implicitly perform linear regression in higher dimensional spaces*

Equivalent Forms of Ridge Regression

- Consider the formulation in which we limit the weights of the coefficients by putting a constraint on size of the L2 norm of the weight vector:

$$\text{argmin}_{\mathbf{w}}(\mathbf{\Phi w} - \mathbf{y})^T(\mathbf{\Phi w} - \mathbf{y})$$
$$\|\mathbf{w}\|_2^2 \leq \xi$$

- The objective function, namely $f(\mathbf{w}) = (\mathbf{\Phi w} - \mathbf{y})^{\mathbf{T}}(\mathbf{\Phi w} - \mathbf{y})$ is strictly convex. The constraint function, $g(\mathbf{w}) = \|\mathbf{w}\|_2^2 - \xi$, is also convex.
- For convex $g(\mathbf{w})$, the set $\{\mathbf{w}|\mathbf{g}(\mathbf{w}) \leq \mathbf{0}\}$, is also convex. (Why?)

Equivalent Forms of Ridge Regression

- To minimize the error function subject to constraint $|\mathbf{w}| \leq \xi$, we apply KKT conditions at the point of optimality $\mathbf{w}^*$

$$\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda \mathbf{g}(\mathbf{w})) = \mathbf{0}$$

(the first KKT condition). Here, $f(\mathbf{w}) = (\Phi\mathbf{w} - \mathbf{y})^T(\Phi\mathbf{w} - \mathbf{y})$ and, $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$.

- Solving we get,

$$\mathbf{w}^* = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}$$

From the second KKT condition we get,

$$\|\mathbf{w}^*\|^2 \leq \xi$$

From the third KKT condition,

$$\lambda \geq 0$$

From the fourth condition

$$\lambda\|\mathbf{w}^*\|^{\mathbf{2}} = \lambda\xi$$

Equivalent Forms of Ridge Regression

- Values of $\mathbf{w}$ and λ that satisfy all these equations would yield an optimal solution. That is, if

$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}\| \leq \xi$$

then $\lambda = 0$ is the solution. Else, for some sufficiently large value, λ will be the solution to

$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}\| = \xi$$

Bound on λ in the regularized least square solution

- Consider,
$$(\Phi^T\Phi + \lambda I)^{-1}\Phi^T \mathbf{y} = \mathbf{w}^*$$
We multiply $(\Phi^T\Phi + \lambda I)$ on both sides and obtain,
$$\|(\Phi^T\Phi)\mathbf{w}^* + (\lambda \mathbf{I})\mathbf{w}^*\| = \|\mathbf{\Phi}^\mathbf{T}\mathbf{y}\|$$
Using the triangle inequality we obtain,
$$\|(\Phi^T\Phi)\mathbf{w}^*\| + (\lambda)\|\mathbf{w}^*\| \geq \|(\mathbf{\Phi}^\mathbf{T}\mathbf{\Phi})\mathbf{w}^* + (\lambda \mathbf{I})\mathbf{w}^*\| = \|\mathbf{\Phi}^\mathbf{T}\mathbf{y}\|$$
- By the Cauchy Shwarz inequality, $\|(\Phi^T\Phi)\mathbf{w}^*\| \leq \alpha\|\mathbf{w}^*\|$ for some $\alpha = \|(\Phi^T\Phi)\|$. Substituting in the previous equation,
$$(\alpha + \lambda)\|\mathbf{w}^*\| \geq \|\mathbf{\Phi}^\mathbf{T}\mathbf{y}\|$$
i.e.
$$\lambda \geq \frac{\|\Phi^T \mathbf{y}\|}{\|\mathbf{w}^*\|} - \alpha$$
Note that when $\|\mathbf{w}^*\| \to \mathbf{0}, \lambda \to \infty$. (Any intuition?) Using $\|\mathbf{w}^*\|^\mathbf{2} \leq \xi$ we get,

Bound on λ in the regularized least square solution

$\|(\Phi^T\Phi)\mathbf{w}^*\| \leq \alpha\|\mathbf{w}^*\|$ for some α for finite $|(\Phi^T\Phi)\mathbf{w}^*\|$. Substituting in the previous equation,

$$(\alpha+\lambda)\|\mathbf{w}^*\| \geq \|\Phi^T\mathbf{y}\|$$

i.e.

$$\lambda \geq \frac{\|\Phi^T\mathbf{y}\|}{\|\mathbf{w}^*\|} - \alpha$$

Note that when $\|\mathbf{w}^*\| \to 0, \lambda \to \infty$. (Any intuition?) Using $\|\mathbf{w}^*\|^2 \leq \xi$ we get,

$$\lambda \geq \frac{\|\Phi^T\mathbf{y}\|}{\sqrt{\xi}} - \alpha$$

This is not the exact solution of λ but the bound proves the existence of λ for some ξ and Φ.

The Resultant alternative objective function

Substituting $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$, in the first KKT equation considered earlier:

$$\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda \cdot (\|\mathbf{w}\|^2 - \xi)) = \mathbf{0}$$

This is equivalent to solving

$$\min(\| \Phi\mathbf{w} - \mathbf{y} \|^2 + \lambda \| \mathbf{w} \|^2)$$

for the same choice of λ. This form of **regularized** ridge regression is the **penalized ridge regression**.