Convex Optimization

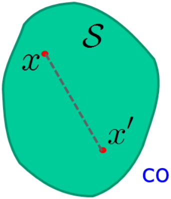
Dani Yogatama

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

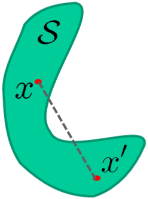
February 12, 2014

Key Concepts in Convex Analysis: Convex Sets

$\mathcal{S}$ is convex if $\quad x, x' \in \mathcal{S} \;\Rightarrow \forall \lambda \in [0,1] \;\; \lambda x + (1-\lambda)x' \in \mathcal{S}$

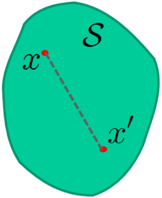


convex

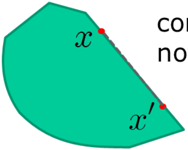


non-convex

$\mathcal{S}$ is strictly convex if $x, x' \in \mathcal{S} \;\Rightarrow \forall \lambda \in (0,1) \;\; \lambda x + (1-\lambda)x' \in \mathrm{int}(\mathcal{S})$



strictly
convex



convex, but
not strictly

Key Concepts in Convex Analysis: Convex Functions

Extended real valued function: $f : \mathbb{R}^N \rightarrow \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$

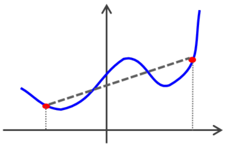
Domain of a function: $\mathrm{dom}(f) = \{x : f(x) \neq +\infty\}$

f is a convex function if

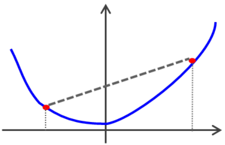
$$\forall \lambda \in [0,1], x, x' \in \mathrm{dom}(f) \;\; f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

f is a strictly convex function if

$$\forall \lambda \in (0,1), x, x' \in \mathrm{dom}(f) \;\; f(\lambda x + (1-\lambda)x') < \lambda f(x) + (1-\lambda) f(x')$$

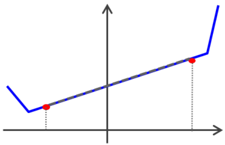


non-convex



convex

strictly convex



convex, not strictly

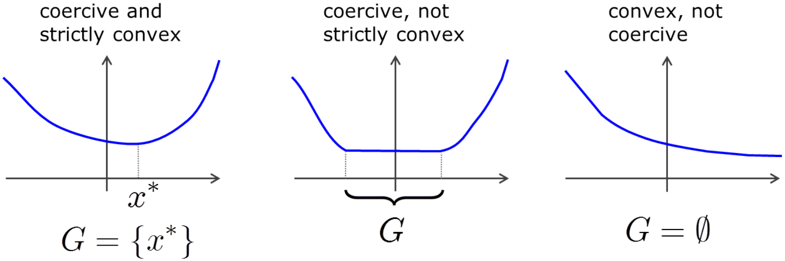
Key Concepts in Convex Analysis: Minimizers

$$f : \mathbb{R}^N \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$$

f is coercive if $\displaystyle\lim_{\|x\| \to +\infty} f(x) = +\infty$

if f is coercive, then $G \equiv \arg\min_x f(x)$ is a non-empty set

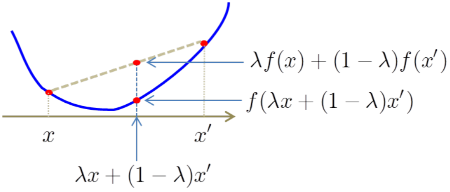
if f is strictly convex, then G has at most one element



Key Concepts in Convex Analysis: Strong Convexity

Recall the definition of convex function: $\forall \lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x')$$



convexity

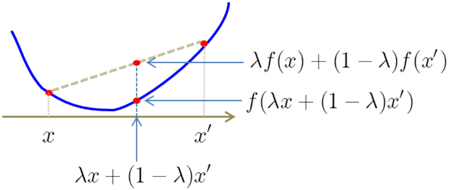
Key Concepts in Convex Analysis: Strong Convexity

Recall the definition of convex function: $\forall \lambda \in [0, 1]$,

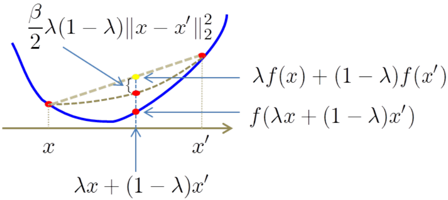
$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda)f(x')$$

A $\beta-$strongly convex function satisfies a stronger condition: $\forall \lambda \in [0, 1]$

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda)f(x') - \frac{\beta}{2}\lambda(1-\lambda)\|x - x'\|_2^2$$



convexity



strong convexity

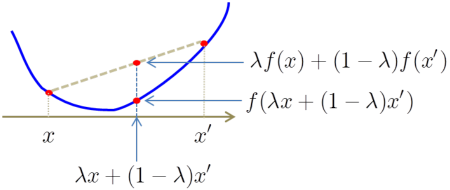
Key Concepts in Convex Analysis: Strong Convexity

Recall the definition of convex function: $\forall \lambda \in [0, 1]$,

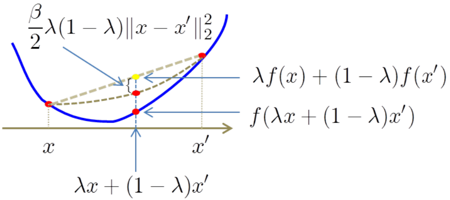
$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

A $\beta-$strongly convex function satisfies a stronger condition: $\forall \lambda \in [0, 1]$

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x') - \frac{\beta}{2}\lambda(1-\lambda)\|x - x'\|_2^2$$



convexity



strong convexity

Strong convexity $\overset{\Rightarrow}{\not\Leftarrow}$ strict convexity.

Key Concepts in Convex Analysis: Subgradients

Convexity $\Rightarrow$ continuity; convexity $\not\Rightarrow$ differentiability (e.g., $f(\mathbf{w}) = \|w\|_1$).

Subgradients generalize gradients for (maybe non-diff.) convex functions:

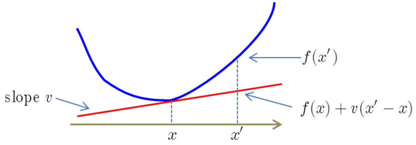
Key Concepts in Convex Analysis: Subgradients

Convexity $\Rightarrow$ continuity; convexity $\not\Rightarrow$ differentiability (e.g., $f(\mathbf{w}) = \|w\|_1$).

Subgradients generalize gradients for (maybe non-diff.) convex functions:

$$\boxed{\mathbf{v} \text{ is a } \textcolor{blue}{\text{subgradient}} \text{ of } f \text{ at } \mathbf{x} \text{ if } \; f(\mathbf{x}') \geq f(\mathbf{x}) + \mathbf{v}^\top(\mathbf{x}' - \mathbf{x})}$$

Subdifferential: $\partial f(\mathbf{x}) = \{\mathbf{v} : \; \mathbf{v} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$



linear lower bound

Key Concepts in Convex Analysis: Subgradients

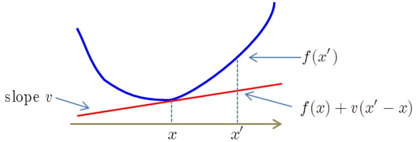
Convexity $\Rightarrow$ continuity; convexity $\not\Rightarrow$ differentiability (e.g., $f(\mathbf{w}) = \|w\|_1$).

Subgradients generalize gradients for (maybe non-diff.) convex functions:

$$\boxed{\mathbf{v} \text{ is a subgradient of } f \text{ at } \mathbf{x} \text{ if } f(\mathbf{x}') \geq f(\mathbf{x}) + \mathbf{v}^\top(\mathbf{x}' - \mathbf{x})}$$

Subdifferential: $\partial f(\mathbf{x}) = \{\mathbf{v} : \mathbf{v} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$

If f is differentiable, $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$



linear lower bound

Key Concepts in Convex Analysis: Subgradients

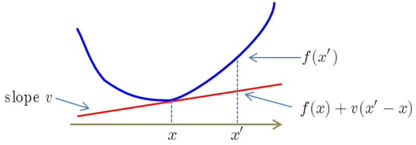
Convexity $\Rightarrow$ continuity; convexity $\not\Rightarrow$ differentiability (e.g., $f(\mathbf{w}) = \|w\|_1$).

Subgradients generalize gradients for (maybe non-diff.) convex functions:

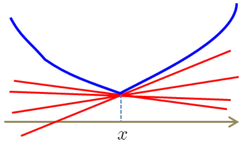
$$\boxed{\mathbf{v} \text{ is a subgradient of } f \text{ at } \mathbf{x} \text{ if } f(\mathbf{x}') \geq f(\mathbf{x}) + \mathbf{v}^\top(\mathbf{x}' - \mathbf{x})}$$

Subdifferential: $\partial f(\mathbf{x}) = \{\mathbf{v} : \mathbf{v} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$

If f is differentiable, $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$



linear lower bound



non-differentiable case

Key Concepts in Convex Analysis: Subgradients

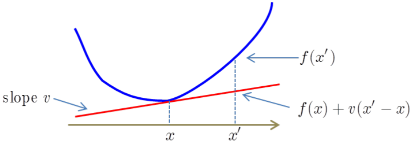
Convexity $\Rightarrow$ continuity; convexity $\not\Rightarrow$ differentiability (e.g., $f(\mathbf{w}) = \|w\|_1$).

Subgradients generalize gradients for (maybe non-diff.) convex functions:

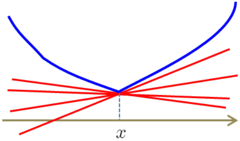
$$\boxed{\mathbf{v} \text{ is a subgradient of } f \text{ at } \mathbf{x} \text{ if } f(\mathbf{x}') \geq f(\mathbf{x}) + \mathbf{v}^\top(\mathbf{x}' - \mathbf{x})}$$

Subdifferential: $\partial f(\mathbf{x}) = \{\mathbf{v} : \mathbf{v} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$

If f is differentiable, $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$



linear lower bound



non-differentiable case

Notation: $\tilde{\nabla} f(\mathbf{x})$ is a subgradient of f at $\mathbf{x}$

Establishing convexity

How to check if $f(x)$ is a convex function?

Establishing convexity

How to check if $f(x)$ is a convex function?

- Verify definition of a convex function.
- Check if $\frac{\partial^2 f(x)}{\partial^2 x}$ greater than or equal to 0 (for twice differentiable function).
- Show that it is constructed from simple convex functions with operations that preserver convexity.
  - nonnegative weighted sum
  - composition with affine function
  - pointwise maximum and supremum
  - composition
  - minimization
  - perspective

Reference: Boyd and Vandenberghe (2004)

Unconstrained Optimization

Algorithms:

- First order methods (gradient descent, FISTA, etc.)
- Higher order methods (Newton's method, ellipsoid, etc.)
- ...

Gradient descent

Problem:

$$\boxed{\min_x f(x)}$$

Gradient descent

Problem:

$$\boxed{\min_x f(x)}$$

Algorithm:

- $g_t = \frac{\partial f(x_t)}{\partial x}$.
- $x_t = x_{t-1} - \eta g_t$.
- Repeat until convergence.

Newton's method

Problem:

$$\boxed{\min_x f(x)}$$

Assume f is twice differentiable.

Newton's method

Problem:

$$\boxed{\min_x f(x)}$$

Assume f is twice differentiable.

Algorithm:

- $g_t = \frac{\partial f(x_t)}{\partial x}$.
- $s_t = H^{-1} g_t$, where H is the Hessian.
- $x_t = x_{t-1} - \eta s_t$.
- Repeat until convergence.

Newton's method

Problem:

$$\boxed{\min_x f(x)}$$

Assume f is twice differentiable.

Algorithm:

- $g_t = \frac{\partial f(x_t)}{\partial x}$.
- $s_t = H^{-1} g_t$, where H is the Hessian.
- $x_t = x_{t-1} - \eta s_t$.
- Repeat until convergence.

Newton's method is a special case of steepest descent using Hessian norm.

Duality

Primal problem:

$$\boxed{\begin{array}{rll} & \min_x f(x) & \\ \text{subject to} & g_i(x) \leq 0 & i = 1, \ldots, m \\ & h_i(x) = 0 & i = 1, \ldots, p \end{array}}$$

for $x \in \mathcal{X}$.

Duality

Primal problem:

$$\boxed{\begin{array}{lll} & \min_x f(x) & \\ \text{subject to} & g_i(x) \leq 0 & i = 1, \dots, m \\ & h_i(x) = 0 & i = 1, \dots, p \end{array}}$$

for $x \in \mathcal{X}$.

Lagrangian:

$$\boxed{\mathcal{L}(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)}$$

λ_i and ν_i are Lagrange multipliers.

Duality

Primal problem:

$$\boxed{\begin{array}{lll} & \min_x f(x) & \\ \text{subject to} & g_i(x) \leq 0 & i = 1, \ldots, m \\ & h_i(x) = 0 & i = 1, \ldots, p \end{array}}$$

for $x \in \mathcal{X}$.

Lagrangian:

$$\boxed{\mathcal{L}(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)}$$

λ_i and ν_i are Lagrange multipliers.

Suppose x is feasible and $\lambda \geq 0$, then we get the lower bound:

$$\boxed{f(x) \geq \mathcal{L}(x, \lambda, \nu) \forall x \in \mathcal{X}, \lambda \in \mathbb{R}_{+}^{m}}$$

Duality

Primal optimal: $\boxed{p^* = \min_x \max_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu)}$

Duality

Primal optimal: $\boxed{p^* = \min_x \max_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu)}$

Lagrange dual function: $\boxed{\min_x \mathcal{L}(x, \lambda, \nu)}$

This is a concave function, regardless of whether $f(x)$ convex or not. Can be $-\infty$ for some λ and ν.

Duality

Primal optimal: $\boxed{p^* = \min_x \max_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu)}$

Lagrange dual function: $\boxed{\min_x \mathcal{L}(x, \lambda, \nu)}$

This is a concave function, regardless of whether $f(x)$ convex or not. Can be $-\infty$ for some λ and ν.

Lagrange dual problem: $\boxed{\max_{\lambda, \nu} \mathcal{L}(x, \lambda, \nu) \text{ subject to } \lambda \geq 0}$

Dual feasible: if $\lambda \geq 0$ and $\lambda, \nu \in$ dom $\mathcal{L}(x, \lambda, \nu)$.

Duality

Primal optimal: $\boxed{p^* = \min_x \max_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu)}$

Lagrange dual function: $\boxed{\min_x \mathcal{L}(x, \lambda, \nu)}$

This is a concave function, regardless of whether $f(x)$ convex or not. Can be $-\infty$ for some λ and ν.

Lagrange dual problem: $\boxed{\max_{\lambda, \nu} \mathcal{L}(x, \lambda, \nu) \text{ subject to } \lambda \geq 0}$

Dual feasible: if $\lambda \geq 0$ and $\lambda, \nu \in \text{ dom } \mathcal{L}(x, \lambda, \nu)$.

Dual optimal: $\boxed{d^* = \max_{\lambda \geq 0, \nu} \min_x \mathcal{L}(x, \lambda, \nu)}$

Duality

Weak duality $p^* \geq d^*$ always holds for convex and nonconvex problems

Duality

Weak duality $p^* \geq d^*$ always holds for convex and nonconvex problems

Strong duality $p^* = d^*$ does not hold in general, but usually holds for convex problems. Strong duality is guaranteed by Slater's constraint qualification.

Strong duality holds if the problem is strictly feasible, i.e.

$$\boxed{\exists x \in \mathtt{int}\,\mathcal{D} \text{ s.t. } g_i(x) < 0, i = 1, \ldots, m, h_i(x) = 0, i = 1, \ldots, p}$$

Duality

Weak duality $p^* \geq d^*$ always holds for convex and nonconvex problems

Strong duality $p^* = d^*$ does not hold in general, but usually holds for convex problems. Strong duality is guaranteed by Slater's constraint qualification.

Strong duality holds if the problem is strictly feasible, i.e.

$$\boxed{\exists x \in \mathtt{int}\,\mathcal{D} \text{ s.t. } g_i(x) < 0, i = 1, \ldots, m, h_i(x) = 0, i = 1, \ldots, p}$$

Assume strong duality holds and p^* and d^* are attained.

$$\boxed{p^* = f(x^*) = d^* = \min_x \mathcal{L}(x^*, \lambda^*, \nu^*) \leq \mathcal{L}(x^*, \lambda^*, \nu^*) \leq f(x^*) = p^*}$$

Duality

Weak duality $p^* \geq d^*$ always holds for convex and nonconvex problems

Strong duality $p^* = d^*$ does not hold in general, but usually holds for convex problems. Strong duality is guaranteed by Slater's constraint qualification.

Strong duality holds if the problem is strictly feasible, i.e.

$$\boxed{\exists x \in \mathtt{int}\,\mathcal{D} \text{ s.t. } g_i(x) < 0, i = 1, \ldots, m, h_i(x) = 0, i = 1, \ldots, p}$$

Assume strong duality holds and p^* and d^* are attained.

$$\boxed{p^* = f(x^*) = d^* = \min_x \mathcal{L}(x^*, \lambda^*, \nu^*) \leq \mathcal{L}(x^*, \lambda^*, \nu^*) \leq f(x^*) = p^*}$$

We have:

- $x^* \in \arg\min_x \mathcal{L}(x^*, \lambda^*, \nu^*)$.
- $\lambda_i^* g_i(x^*) = 0$ for $i = 1, \ldots, m$ (complementary slackness).

Karush-Kuhn-Tucker condition

For a differentiable $g(x)$ and $h(x)$, the KKT conditions are:

$$
\begin{aligned}
& g_i(x^*) \leq 0, h_i(x^*) = 0, && \text{primal feasibility} \\
& \lambda_i^* \geq 0, && \text{dual feasibility} \\
& \lambda_i^* g_i(x^*) = 0, && \text{complementary slackness} \\
& \frac{\partial \mathcal{L}(x^*, \lambda^*, \nu^*)}{\partial x}\Big|_{x=x^*} = 0 && \text{Lagrangian stationarity}
\end{aligned}
$$

Karush-Kuhn-Tucker condition

For a differentiable $g(x)$ and $h(x)$, the KKT conditions are:

$$\begin{aligned}
&g_i(x^*) \leq 0, h_i(x^*) = 0, && \text{primal feasibility} \\
&\lambda_i^* \geq 0, && \text{dual feasibility} \\
&\lambda_i^* g_i(x^*) = 0, && \text{complementary slackness} \\
&\frac{\partial \mathcal{L}(x^*, \lambda^*, \nu^*)}{\partial x}\Big|_{x=x^*} = 0 && \text{Lagrangian stationarity}
\end{aligned}$$

If $\hat{x}, \hat{\lambda}, \hat{\nu}$ satify the KKT for a convex problem, they are optimal.

Support Vector Machines

Primal problem (hard constraint):

$$\begin{aligned} \min_w \quad & \frac{1}{2}\|w\|_2^2 \\ \text{subject to} \quad & y_i \langle x_i, w \rangle \geq 1, i = 1, \ldots, n \end{aligned}$$

Support Vector Machines

Primal problem (hard constraint):

$$\begin{array}{ll} \min_w & \frac{1}{2}\|w\|_2^2 \\ \text{subject to} & y_i\langle x_i, w\rangle \geq 1, i = 1, \dots, n \end{array}$$

Lagrangian:

$$\mathcal{L}(w, \lambda) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \lambda_i (y_i \langle x_i, w\rangle - 1)$$

Support Vector Machines

Primal problem (hard constraint):

$$\begin{aligned} \min_w \quad & \frac{1}{2}\|w\|_2^2 \\ \text{subject to} \quad & y_i\langle x_i, w\rangle \geq 1, i = 1, \ldots, n \end{aligned}$$

Lagrangian:

$$\mathcal{L}(w, \lambda) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \lambda_i(y_i\langle x_i, w\rangle - 1)$$

Minimizing with respect to w, we have:

$$\begin{aligned} \tfrac{\partial \mathcal{L}(w,\lambda)}{\partial w} &= 0 \\ w - \textstyle\sum_{i=1}^{n} \lambda_i y_i x_i &= 0 \\ w &= \sum_{i=1}^{n} \lambda_i y_i x_i \end{aligned}$$

Support Vector Machines

Plug this back into the Lagrangian:

$$\mathcal{L}(\lambda) = \sum_{1=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^\top x_j$$

Support Vector Machines

Plug this back into the Lagrangian:

$$\mathcal{L}(\lambda) = \sum_{1=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^\top x_j$$

Lagrange dual problem is:

$$\begin{aligned} \max_{\lambda} \quad & \sum_{1=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^\top x_j \\ \text{subject to} \quad & \lambda_i \geq 0, i = 1, \ldots, n \\ & \sum_{i=1}^{n} \lambda_i y_i = 0 \end{aligned}$$

Support Vector Machines

Plug this back into the Lagrangian:

$$\mathcal{L}(\lambda) = \sum_{1=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^\top x_j$$

Lagrange dual problem is:

$$\begin{aligned} \max_\lambda \quad & \sum_{1=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^\top x_j \\ \text{subject to} \quad & \lambda_i \geq 0, i = 1, \ldots, n \\ & \sum_{i=1}^{n} \lambda_i y_i = 0 \end{aligned}$$

Since this problem only depends on $x_i^\top x_j$, we can use kernels and learn in high dimensional space without having to explicitly represent $\phi(x)$.

Support Vector Machines

Primal problem (soft constraint):

$$
\begin{aligned}
\min_w \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \\
\text{subject to} \quad & y_i\langle x_i, w\rangle \geq 1 - \xi_i, i = 1, \ldots, n \\
& \xi_i \geq 0, i = 1\ldots, n
\end{aligned}
$$

Support Vector Machines

Lagrange dual problem for the soft constraint:

$$\begin{aligned}
\max_{\lambda} \quad & \sum_{1=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^\top x_j & (1) \\
\text{subject to} \quad & 0 \leq \lambda_i \leq C, i = 1, \ldots, n & (2) \\
& \sum_{i=1}^{n} \lambda_i y_i = 0 & (3)
\end{aligned}$$

Support Vector Machines

Lagrange dual problem for the soft constraint:

$$\max_\lambda \quad \sum_{1=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^\top x_j \tag{1}$$

$$\text{subject to} \quad 0 \le \lambda_i \le C, i = 1, \ldots, n \tag{2}$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0 \tag{3}$$

KKT conditions, for all i:

$$\lambda_i = 0 \quad \rightarrow y_i \langle x_i, w \rangle \ge 1 \tag{4}$$

$$0 < \lambda_i < C \quad \rightarrow y_i \langle x_i, w \rangle = 1 \tag{5}$$

$$\lambda_i = C \quad \rightarrow y_i \langle x_i, w \rangle \le 1 \tag{6}$$

Sequential Minimal Optimization (Platt, 1998)

An efficient way to solve SVM dual problem. Break a large QP program into a series of smallest possible QP problems. Solve these small subproblems analytically.

Sequential Minimal Optimization (Platt, 1998)

An efficient way to solve SVM dual problem. Break a large QP program into a series of smallest possible QP problems. Solve these small subproblems analytically.

In a nutshell

- Choose two Lagrange multipliers λ_i and λ_j.
- Optimize the dual problem with respect to these two Lagrange multipliers, holding others fixed.
- Repeat until convergence.

Sequential Minimal Optimization (Platt, 1998)

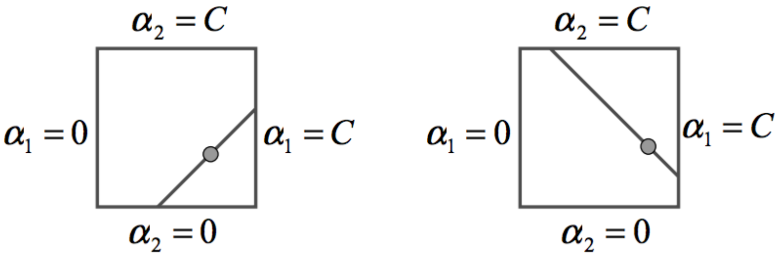
An efficient way to solve SVM dual problem. Break a large QP program into a series of smallest possible QP problems. Solve these small subproblems analytically.

In a nutshell

- Choose two Lagrange multipliers λ_i and λ_j.
- Optimize the dual problem with respect to these two Lagrange multipliers, holding others fixed.
- Repeat until convergence.

There are heuristics to choose Lagrange multipliers that maximizes the step size towards the global maximum. The first one is chosen from examples that violate the KKT condition. The second one is chosen using approximation based on absolute difference in error values (see Platt (1998)).

Sequential Minimal Optimization (Platt, 1998)



$y_1 \neq y_2 \Rightarrow \alpha_1 - \alpha_2 = \gamma$ $\qquad$ $y_1 = y_2 \Rightarrow \alpha_1 + \alpha_2 = \gamma$

Sequential Minimal Optimization (Platt, 1998)

For any two Lagrange multipliers, the constraints are::

$$0 < \lambda_i, \lambda_j < C \tag{7}$$

$$y_i\lambda_i + y_j\lambda_j = -\sum_{k \neq i,j} y_k\lambda_k = \gamma \tag{8}$$

Sequential Minimal Optimization (Platt, 1998)

For any two Lagrange multipliers, the constraints are::

$$0 < \lambda_i, \lambda_j < C \tag{7}$$

$$y_i\lambda_i + y_j\lambda_j = -\sum_{k \neq i,j} y_k\lambda_k = \gamma \tag{8}$$

Express λ_i in terms of λ_j

$$\lambda_i = \frac{\gamma - \lambda_j y_j}{y_i}$$

Sequential Minimal Optimization (Platt, 1998)

For any two Lagrange multipliers, the constraints are::

$$0 < \lambda_i, \lambda_j < C \tag{7}$$

$$y_i\lambda_i + y_j\lambda_j = -\sum_{k\neq i,j} y_k\lambda_k = \gamma \tag{8}$$

Express λ_i in terms of λ_j

$$\lambda_i = \frac{\gamma - \lambda_j y_j}{y_i}$$

Plug this back into our original function. We are then left with a very simple quadratic problem with one variable λ_j

Sequential Minimal Optimization (Platt, 1998)

Solve for the second Lagrange multiplier λ_j.

Sequential Minimal Optimization (Platt, 1998)

Solve for the second Lagrange multiplier λ_j.

If $y_i \neq y_j$, the following bounds apply to λ_j:

$$
\begin{aligned}
L &= \max(0, \lambda_j^{t-1} - \lambda_i^{y-1}) && (9)\\
H &= \min(C, C + \lambda_j^{t-1} - \lambda_i^{y-1}) && (10)
\end{aligned}
$$

If $y_i = y_j$, the following bounds apply to λ_j:

$$
\begin{aligned}
L &= \max(0, \lambda_j^{t-1} + \lambda_i^{y-1} - C) && (11)\\
H &= \min(C, \lambda_j^{t-1} + \lambda_i^{y-1}) && (12)
\end{aligned}
$$

Sequential Minimal Optimization (Platt, 1998)

Solve for the second Lagrange multiplier λ_j.

If $y_i \neq y_j$, the following bounds apply to λ_j:

$$L = \max(0, \lambda_j^{t-1} - \lambda_i^{y-1}) \tag{9}$$

$$H = \min(C, C + \lambda_j^{t-1} - \lambda_i^{y-1}) \tag{10}$$

If $y_i = y_j$, the following bounds apply to λ_j:

$$L = \max(0, \lambda_j^{t-1} + \lambda_i^{y-1} - C) \tag{11}$$

$$H = \min(C, \lambda_j^{t-1} + \lambda_i^{y-1}) \tag{12}$$

The solution is:

$$\lambda_j = \begin{cases} H & \text{if} \lambda_j > H \\ \lambda_j & \text{if} L \leq \lambda_j \leq H \\ L & \text{if} \lambda_j < L \end{cases}$$

Fenchel duality

If a convex conjugate of $f(x)$ is known, the dual function can be easily derived. The convex conjugate of a function f is:

$$f^*(y) = \max_x \langle y, x \rangle - f(x) \tag{13}$$

Fenchel duality

If a convex conjugate of $f(x)$ is known, the dual function can be easily derived. The convex conjugate of a function f is:

$$f^*(y) = \max_x \langle y, x \rangle - f(x) \tag{13}$$

For a generic problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & Ax \leq b \\ & Cx = d \end{aligned}$$

The dual function is: $-f^*(-A^\top \lambda - C^\top \nu) - b^\top \lambda - d^\top \nu$

Fenchel duality

If a convex conjugate of $f(x)$ is known, the dual function can be easily derived. The convex conjugate of a function f is:

$$f^*(y) = \max_x \langle y, x \rangle - f(x) \tag{13}$$

For a generic problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & Ax \leq b \\ & Cx = d \end{aligned}$$

The dual function is: $-f^*(-A^\top \lambda - C^\top \nu) - b^\top \lambda - d^\top \nu$

There are many functions whose conjugate are easy to compute:

- Exponential
- Logistic
- Quadratic form
- Log determinant
- ...

Parting notes

Dual formulation is useful.

- Give new insights into our problem,
- Allow us to develop better optimization methods and use kernel tricks.

Thank you!

- Questions?

References I

Some slides are from an upcoming EACL 2014 tutorial with Andre F. T. Martins, Noah A. Smith, and Mario F. T. Figueiredo

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In Scholkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.