Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2018

L5: Optimization Dual

Goals of this lecture

Goals of this lecture

- Constrained optimization

Goals of this lecture

- Constrained optimization
- Understanding duality for optimization

Goals of this lecture

- Constrained optimization
- Understanding duality for optimization

Reading Material

Goals of this lecture

- Constrained optimization
- Understanding duality for optimization

Reading Material

- S. Boyd and L. Vandenberghe; Convex Optimization; Chapter 5

Optimization problems that we have seen so far:

Optimization problems that we have seen so far:

- Linear Regression

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w}\right)^2$$

Optimization problems that we have seen so far:

- Linear Regression

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w}\right)^2$$

- Logistic Regression

$$\min_{\boldsymbol{w}} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \boldsymbol{w}^T \phi(x^{(i)}))\right)$$

Optimization problems that we have seen so far:

- Linear Regression

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w}\right)^2$$

- Logistic Regression

$$\min_{\boldsymbol{w}} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)}\boldsymbol{w}^T\phi(x^{(i)}))\right)$$

Finding optimum:

Optimization problems that we have seen so far:

- Linear Regression

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w}\right)^2$$

- Logistic Regression

$$\min_{\boldsymbol{w}} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \boldsymbol{w}^T \phi(x^{(i)}))\right)$$

Finding optimum:

Analytically computable optimum vs. gradient descent

The Problem more generally:

The Problem more generally:

$$\begin{array}{ll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 \quad \forall i \in \{1, \ldots, C\} \end{array}$$

The Problem more generally:

$$\begin{array}{ll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 \quad \forall i \in \{1, \ldots, C\} \end{array}$$

Solution:

The Problem more generally:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & f_0(\boldsymbol{w}) \\ \text{s.t.} \quad & f_i(\boldsymbol{w}) \leq 0 \quad \forall i \in \{1, \ldots, C\} \end{aligned}$$

Solution:

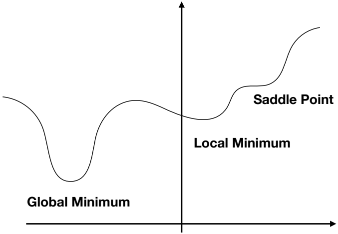
Solution $\boldsymbol{w}^*$ has smallest value $f_0(\boldsymbol{w}^*)$ among all values that satisfy constraints

The Problem more generally:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & f_0(\boldsymbol{w}) \\ \text{s.t.} \quad & f_i(\boldsymbol{w}) \leq 0 \quad \forall i \in \{1, \ldots, C\} \end{aligned}$$

Solution:

Solution $\boldsymbol{w}^*$ has smallest value $f_0(\boldsymbol{w}^*)$ among all values that satisfy constraints



Original/Primal Problem:

Original/Primal Problem:

$$\begin{array}{lll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) & \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 & \forall i \in \{1, \ldots, C_1\} \\ & h_i(\boldsymbol{w}) = 0 & \forall i \in \{1, \ldots, C_2\} \end{array}$$

Original/Primal Problem:

$$\begin{array}{lll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) & \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 & \forall i \in \{1, \ldots, C_1\} \\ & h_i(\boldsymbol{w}) = 0 & \forall i \in \{1, \ldots, C_2\} \end{array}$$

How to optimize this?

When can we find the optimum?

When can we find the optimum?

- Least squares, linear and convex programs can be solved efficiently and reliably

When can we find the optimum?

- Least squares, linear and convex programs can be solved efficiently and reliably
- General optimization problems are very difficult to solve

When can we find the optimum?

- Least squares, linear and convex programs can be solved efficiently and reliably
- General optimization problems are very difficult to solve
- Often compromise between accuracy and computation time

What's the form of 'least squares,' 'linear,' and 'convex' programs?

- Least squares program

What's the form of 'least squares,' 'linear,' and 'convex' programs?

- Least squares program

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w}\right)^2$$

What's the form of 'least squares,' 'linear,' and 'convex' programs?

- Least squares program

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w}\right)^2$$

- Linear program

What's the form of 'least squares,' 'linear,' and 'convex' programs?

- Least squares program

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w} \right)^2$$

- Linear program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

What's the form of 'least squares,' 'linear,' and 'convex' programs?

- Least squares program

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w}\right)^2$$

- Linear program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

- Convex program

What's the form of 'least squares,' 'linear,' and 'convex' programs?

- Least squares program

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \boldsymbol{w}\right)^2$$

- Linear program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

- Convex program when all f_i convex (generalizes the above)

$$\min_{\boldsymbol{w}} f_0(\boldsymbol{w}) \quad \text{s.t.} \quad f_i(\boldsymbol{w}) \leq 0 \quad \forall i \in \{1, \ldots, C\}$$

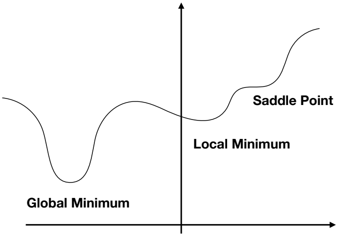
Optimality of convex optimization

Optimality of convex optimization

- A point $\boldsymbol{w}^*$ is locally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})\ \forall \boldsymbol{w}$ in a neighborhood of $\boldsymbol{w}^*$; globally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})\ \forall \boldsymbol{w}$

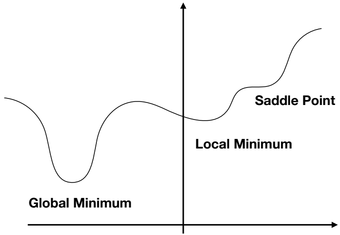
Optimality of convex optimization

- A point $\boldsymbol{w}^*$ is locally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w}) \; \forall \boldsymbol{w}$ in a neighborhood of $\boldsymbol{w}^*$; globally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w}) \; \forall \boldsymbol{w}$



Optimality of convex optimization

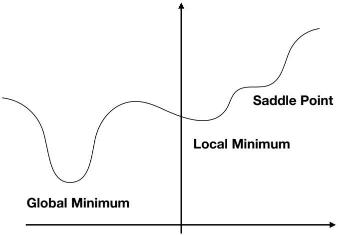
- A point $\boldsymbol{w}^*$ is locally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})\ \forall \boldsymbol{w}$ in a neighborhood of $\boldsymbol{w}^*$; globally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})\ \forall \boldsymbol{w}$



For convex problems global optimality follows directly from local optimality.

Optimality of convex optimization

- A point $\boldsymbol{w}^*$ is locally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})$ $\forall \boldsymbol{w}$ in a neighborhood of $\boldsymbol{w}^*$; globally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})$ $\forall \boldsymbol{w}$

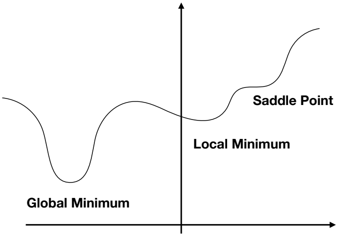


For convex problems global optimality follows directly from local optimality.

- For a local minimum of f, $\nabla f(\boldsymbol{w}^*) = 0$

Optimality of convex optimization

- A point $\boldsymbol{w}^*$ is locally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})\ \forall \boldsymbol{w}$ in a neighborhood of $\boldsymbol{w}^*$; globally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})\ \forall \boldsymbol{w}$

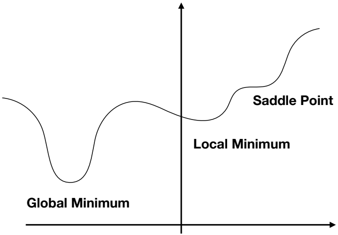


For convex problems global optimality follows directly from local optimality.

- For a local minimum of f, $\nabla f(\boldsymbol{w}^*) = 0$
- If f convex, then $\nabla f(\boldsymbol{w}^*) = 0$ sufficient for global optimality

Optimality of convex optimization

- A point $\boldsymbol{w}^*$ is locally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})\ \forall \boldsymbol{w}$ in a neighborhood of $\boldsymbol{w}^*$; globally optimal if $f(\boldsymbol{w}^*) \leq f(\boldsymbol{w})\ \forall \boldsymbol{w}$



For convex problems global optimality follows directly from local optimality.

- For a local minimum of f, $\nabla f(\boldsymbol{w}^*) = 0$
- If f convex, then $\nabla f(\boldsymbol{w}^*) = 0$ sufficient for global optimality

This makes convex optimization special!

Algorithms to search for the optimum?

Descent methods

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

Intuition

Descent methods

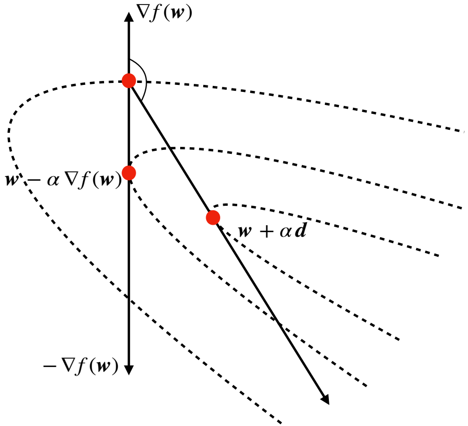
$$\min_{\boldsymbol{w}} f(\boldsymbol{w})$$

Intuition (find a stationary point with $\nabla f(\boldsymbol{w}) = 0$)

Descent methods

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

Intuition (find a stationary point with $\nabla f(\boldsymbol{w}) = 0$)



Iterative algorithm

Iterative algorithm

- Start with some guess $\boldsymbol{w}$

Iterative algorithm

- Start with some guess $\boldsymbol{w}$
- Iterate k = 1, 2, 3, ...

Iterative algorithm

- Start with some guess $\boldsymbol{w}$
- Iterate k = 1, 2, 3, ...
  - Select direction $\boldsymbol{d}_k$ and stepsize α_k

Iterative algorithm

- Start with some guess $\boldsymbol{w}$
- Iterate k = 1, 2, 3, ...
  - Select direction $\boldsymbol{d}_k$ and stepsize α_k
  - $\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_k \boldsymbol{d}_k$

Iterative algorithm

- Start with some guess $\boldsymbol{w}$
- Iterate k = 1, 2, 3, ...
  - Select direction $\boldsymbol{d}_k$ and stepsize α_k
  - $\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_k \boldsymbol{d}_k$
  - Check whether we should stop (e.g., if $\nabla f(\boldsymbol{w}) \approx 0$)

Iterative algorithm

- Start with some guess $\boldsymbol{w}$
- Iterate k = 1, 2, 3, ...
  - Select direction $\boldsymbol{d}_k$ and stepsize α_k
  - $\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_k \boldsymbol{d}_k$
  - Check whether we should stop (e.g., if $\nabla f(\boldsymbol{w}) \approx 0$)

Descent direction d_k satisfies

Iterative algorithm

- Start with some guess $\boldsymbol{w}$
- Iterate k = 1, 2, 3, ...
  - Select direction $\boldsymbol{d}_k$ and stepsize α_k
  - $\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_k \boldsymbol{d}_k$
  - Check whether we should stop (e.g., if $\nabla f(\boldsymbol{w}) \approx 0$)

Descent direction d_k satisfies $\nabla f(\boldsymbol{w})^\top \boldsymbol{d}_k < 0$

How to select direction:

How to select direction:

- Steepest descent: $\boldsymbol{d}_k = -\nabla f(\boldsymbol{w}_k)$

How to select direction:

- Steepest descent: $\boldsymbol{d}_k = -\nabla f(\boldsymbol{w}_k)$
- Scaled gradient: $\boldsymbol{d}_k = -\boldsymbol{D}_k \nabla f(\boldsymbol{w}_k)$ for $\boldsymbol{D}_k \succ 0$

How to select direction:

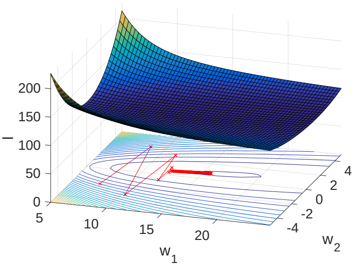
- Steepest descent: $\boldsymbol{d}_k = -\nabla f(\boldsymbol{w}_k)$
- Scaled gradient: $\boldsymbol{d}_k = -\boldsymbol{D}_k \nabla f(\boldsymbol{w}_k)$ for $\boldsymbol{D}_k \succ 0$
  - E.g., Newton's method: $\boldsymbol{D}_k = [\nabla^2 f(\boldsymbol{w}_k)]^{-1}$

How to select direction:

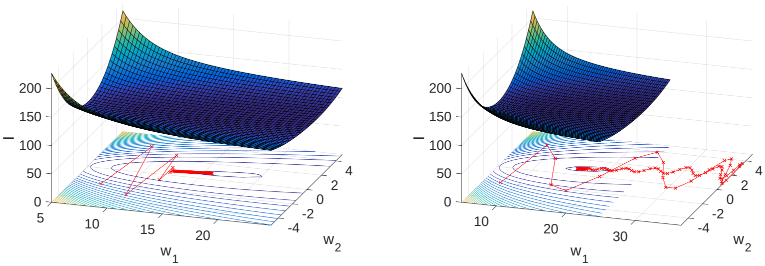
- Steepest descent: $\boldsymbol{d}_k = -\nabla f(\boldsymbol{w}_k)$
- Scaled gradient: $\boldsymbol{d}_k = -\boldsymbol{D}_k \nabla f(\boldsymbol{w}_k)$ for $\boldsymbol{D}_k \succ 0$
  - E.g., Newton's method: $\boldsymbol{D}_k = [\nabla^2 f(\boldsymbol{w}_k)]^{-1}$
- Gradient with momentum

Gradient with momentum

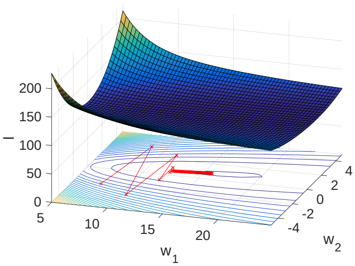
Gradient with momentum
Intuition:

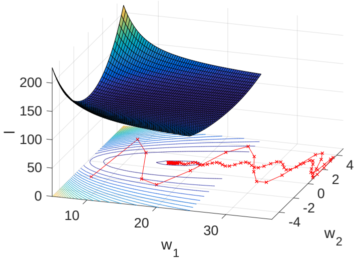


Gradient with momentum
Intuition:



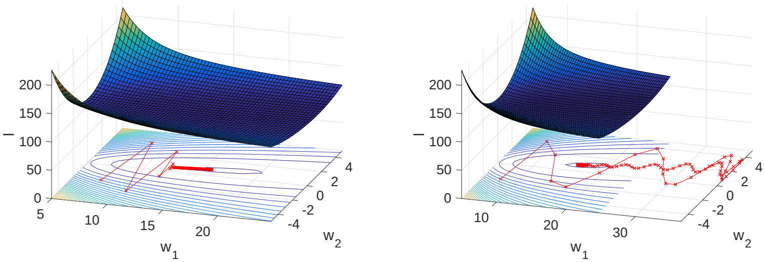
Gradient with momentum
Intuition:





Video

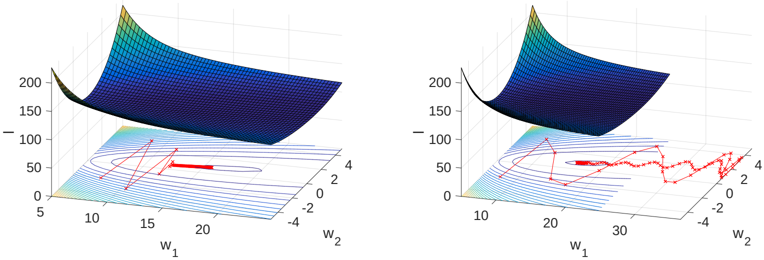
Gradient with momentum
Intuition:



Video

- Polyak's method (aka heavy-ball)

Gradient with momentum
Intuition:

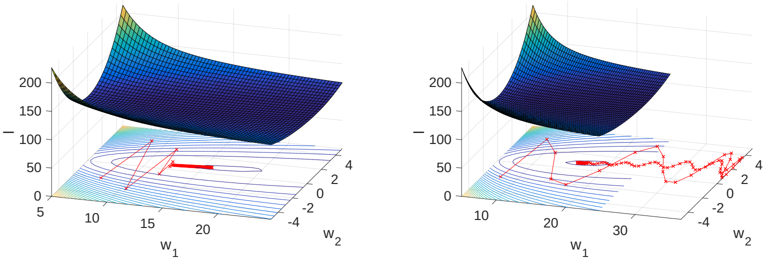


Video

- Polyak's method (aka heavy-ball)

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \nabla f(\mathbf{w}_k) + \beta_k(\mathbf{w}_k - \mathbf{w}_{k-1})$$

Gradient with momentum
Intuition:



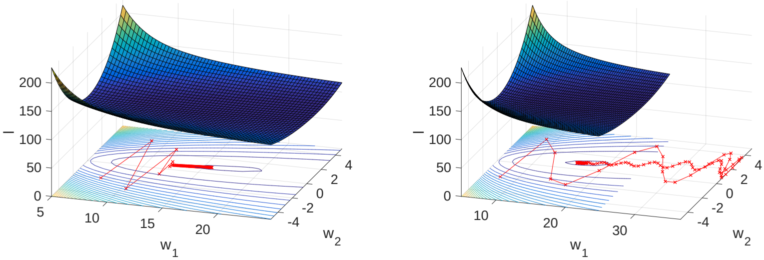
Video

- Polyak's method (aka heavy-ball)

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k) + \beta_k (\boldsymbol{w}_k - \boldsymbol{w}_{k-1})$$

- Momentum method in deep learning

Gradient with momentum
Intuition:



Video

- Polyak's method (aka heavy-ball)

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \nabla f(\mathbf{w}_k) + \beta_k(\mathbf{w}_k - \mathbf{w}_{k-1})$$

- Momentum method in deep learning

$$\begin{aligned} \mathbf{v}_{k+1} &= \beta \mathbf{v}_k + \nabla f(\mathbf{w}_k) \\ \mathbf{w}_{k+1} &= \mathbf{w}_k - \alpha \mathbf{v}_{k+1} \end{aligned}$$

How to select stepsize:

How to select stepsize:

- Exact: $\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{w}_k + \alpha \boldsymbol{d}_k)$

How to select stepsize:

- Exact: $\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{w}_k + \alpha \boldsymbol{d}_k)$
- Constant: $\alpha_k = 1/L$ (for suitable L)

How to select stepsize:

- Exact: $\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{w}_k + \alpha \boldsymbol{d}_k)$
- Constant: $\alpha_k = 1/L$ (for suitable L)
- Diminishing: $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$ (e.g., $\alpha_k = 1/k$)

How to select stepsize:

- Exact: $\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{w}_k + \alpha \boldsymbol{d}_k)$
- Constant: $\alpha_k = 1/L$ (for suitable L)
- Diminishing: $\alpha_k \rightarrow 0$ but $\sum_k \alpha_k = \infty$ (e.g., $\alpha_k = 1/k$)
- Armijo Rule

Recall the structure of our optimization problems:

Recall the structure of our optimization problems:

$$\min_{w} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \ell(y_i, F(x^{(i)}, \boldsymbol{w}))$$

Recall the structure of our optimization problems:

$$\min_{w} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \ell(y_i, F(x^{(i)}, \boldsymbol{w}))$$

- So far we didn't consider the time for computing the gradient

Recall the structure of our optimization problems:

$$\min_{w} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \ell(y_i, F(x^{(i)}, \boldsymbol{w}))$$

- So far we didn't consider the time for computing the gradient
- Iteration complexity is linear in the number of samples $|\mathcal{D}|$

Recall the structure of our optimization problems:

$$\min_{w} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \ell(y_i, F(x^{(i)}, \boldsymbol{w}))$$

- So far we didn't consider the time for computing the gradient
- Iteration complexity is linear in the number of samples $|\mathcal{D}|$
- A large dataset makes gradient computation slow

Recall the structure of our optimization problems:

$$\min_{w} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \ell(y_i, F(x^{(i)}, \boldsymbol{w}))$$

- So far we didn't consider the time for computing the gradient
- Iteration complexity is linear in the number of samples $|\mathcal{D}|$
- A large dataset makes gradient computation slow

How to deal with this?

Stochastic gradient descent

Stochastic gradient descent

Consider a subset of samples and approximate the gradient based on this batch of data.

Stochastic gradient descent

Consider a subset of samples and approximate the gradient based on this batch of data.

- Select a subset of samples $\mathcal{B}_k$

Stochastic gradient descent

Consider a subset of samples and approximate the gradient based on this batch of data.

- Select a subset of samples $\mathcal{B}_k$
- Gradient update using approximation

$$\nabla f(\boldsymbol{w}) \approx \sum_{(x^{(i)},y^{(i)})\in\mathcal{B}_k} \nabla \ell(y^{(i)}, F(x^{(i)}, \boldsymbol{w}))$$

How about constraints?

Original/Primal Problem:

$$\begin{array}{lll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) & \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 & \forall i \in \{1, \ldots, C_1\} \\ & h_i(\boldsymbol{w}) = 0 & \forall i \in \{1, \ldots, C_2\} \end{array}$$

Original/Primal Problem:

$$
\begin{array}{lll}
\min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) & \\
\text{s.t.} & f_i(\boldsymbol{w}) \leq 0 & \forall i \in \{1, \ldots, C_1\} \\
 & h_i(\boldsymbol{w}) = 0 & \forall i \in \{1, \ldots, C_2\}
\end{array}
$$

Lagrangian

Original/Primal Problem:

$$
\begin{array}{lll}
\min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) & \\
\text{s.t.} & f_i(\boldsymbol{w}) \leq 0 & \forall i \in \{1, \ldots, C_1\} \\
 & h_i(\boldsymbol{w}) = 0 & \forall i \in \{1, \ldots, C_2\}
\end{array}
$$

Lagrangian

$$
L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})
$$

Original/Primal Problem:

$$\begin{array}{lll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) & \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 & \forall i \in \{1, \ldots, C_1\} \\ & h_i(\boldsymbol{w}) = 0 & \forall i \in \{1, \ldots, C_2\} \end{array}$$

Lagrangian

$$L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

- λ_i are Lagrange multiplier associated with inequality constraints

Original/Primal Problem:

$$\begin{array}{lll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) & \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 & \forall i \in \{1, \ldots, C_1\} \\ & h_i(\boldsymbol{w}) = 0 & \forall i \in \{1, \ldots, C_2\} \end{array}$$

Lagrangian

$$L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

- λ_i are Lagrange multiplier associated with inequality constraints
- ν_i are Lagrange multiplier associated with equality constraints

Properties of Lagrangian:

$$L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

Properties of Lagrangian:

$$L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

If $\hat{\boldsymbol{w}}$ feasible and $\lambda_i \geq 0 \; \forall i$ then

Properties of Lagrangian:

$$L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

If $\hat{\boldsymbol{w}}$ feasible and $\lambda_i \geq 0\ \forall i$ then

$$f_0(\hat{\boldsymbol{w}}) \geq L(\hat{\boldsymbol{w}}, \lambda, \nu) \geq \min_{\boldsymbol{w}\in\mathcal{W}} L(\boldsymbol{w}, \lambda, \nu) = g(\lambda, \nu) \quad \forall \lambda \geq 0, \nu$$

$$f_0(\boldsymbol{w}^*) \geq g(\lambda, \nu) \quad \forall \lambda \geq 0, \nu$$

$\mathcal{W}$ denotes all the constraints that are not part of the Lagrangian (larger than feasible set)

Properties of Lagrangian:

$$L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

If $\hat{\boldsymbol{w}}$ feasible and $\lambda_i \geq 0 \; \forall i$ then

$$f_0(\hat{\boldsymbol{w}}) \geq L(\hat{\boldsymbol{w}}, \lambda, \nu) \geq \min_{\boldsymbol{w} \in \mathcal{W}} L(\boldsymbol{w}, \lambda, \nu) = g(\lambda, \nu) \quad \forall \lambda \geq 0, \nu$$

$$f_0(\boldsymbol{w}^*) \geq g(\lambda, \nu) \quad \forall \lambda \geq 0, \nu$$

$\mathcal{W}$ denotes all the constraints that are not part of the Lagrangian (larger than feasible set)

Dual Program:

$$\max_{\lambda, \nu} g(\lambda, \nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

Recipe for computing dual program:

Recipe for computing dual program:

- Bring primal program into standard form

Recipe for computing dual program:

- Bring primal program into standard form
- Assign Lagrange multipliers to a suitable set of constraints

Recipe for computing dual program:

- Bring primal program into standard form
- Assign Lagrange multipliers to a suitable set of constraints
- Subsume all other constrains in $\mathcal{W}$

Recipe for computing dual program:

- Bring primal program into standard form
- Assign Lagrange multipliers to a suitable set of constraints
- Subsume all other constrains in $\mathcal{W}$
- Write down the Lagrangian L

Recipe for computing dual program:

- Bring primal program into standard form
- Assign Lagrange multipliers to a suitable set of constraints
- Subsume all other constrains in $\mathcal{W}$
- Write down the Lagrangian L
- Minimize Lagrangian w.r.t. primal variables s.t. $\boldsymbol{w} \in \mathcal{W}$

Examples: Linear Program

Examples: Linear Program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

Examples: Linear Program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

Lagrangian:

Examples: Linear Program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

Lagrangian: $(\lambda \geq 0)$

$$L() = \boldsymbol{c}^\top \boldsymbol{w} + \lambda^\top (\boldsymbol{A}\boldsymbol{w} - \mathbf{b}) = (c + \boldsymbol{A}^\top \lambda)^\top \boldsymbol{w} - \mathbf{b}^\top \lambda$$

Examples: Linear Program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

Lagrangian: $(\lambda \geq 0)$

$$L() = \boldsymbol{c}^\top \boldsymbol{w} + \lambda^\top(\boldsymbol{A}\boldsymbol{w} - \mathbf{b}) = (c + \boldsymbol{A}^\top \lambda)^\top \boldsymbol{w} - \mathbf{b}^\top \lambda$$

Minimizing Lagrangian w.r.t. primal variables:

Examples: Linear Program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

Lagrangian: $(\lambda \geq 0)$

$$L() = \boldsymbol{c}^\top \boldsymbol{w} + \lambda^\top(\boldsymbol{A}\boldsymbol{w} - \mathbf{b}) = (c + \boldsymbol{A}^\top \lambda)^\top \boldsymbol{w} - \mathbf{b}^\top \lambda$$

Minimizing Lagrangian w.r.t. primal variables:

$$\min_{\boldsymbol{w}} L() = \begin{cases} -\mathbf{b}^\top \lambda & \boldsymbol{A}^\top \lambda + \boldsymbol{c} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Examples: Linear Program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

Lagrangian: $(\lambda \geq 0)$

$$L() = \boldsymbol{c}^\top \boldsymbol{w} + \lambda^\top(\boldsymbol{A}\boldsymbol{w} - \mathbf{b}) = (c + \boldsymbol{A}^\top\lambda)^\top \boldsymbol{w} - \mathbf{b}^\top\lambda$$

Minimizing Lagrangian w.r.t. primal variables:

$$\min_{\boldsymbol{w}} L() = \begin{cases} -\mathbf{b}^\top\lambda & \boldsymbol{A}^\top\lambda + \boldsymbol{c} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Dual Program:

Examples: Linear Program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

Lagrangian: $(\lambda \geq 0)$

$$L() = \boldsymbol{c}^\top \boldsymbol{w} + \lambda^\top (\boldsymbol{A}\boldsymbol{w} - \mathbf{b}) = (c + \boldsymbol{A}^\top \lambda)^\top \boldsymbol{w} - \mathbf{b}^\top \lambda$$

Minimizing Lagrangian w.r.t. primal variables:

$$\min_{\boldsymbol{w}} L() = \begin{cases} -\mathbf{b}^\top \lambda & \boldsymbol{A}^\top \lambda + \boldsymbol{c} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Dual Program:

$$\max_{\lambda \geq 0} -\mathbf{b}^\top \lambda \quad \text{s.t.} \quad \boldsymbol{A}^\top \lambda + \boldsymbol{c} = 0,$$

Examples: Logistic Regression

Examples: Logistic Regression

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})))$$

Examples: Logistic Regression

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})))$$

Reformulate:

Examples: Logistic Regression

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})))$$

Reformulate:

$$\min_{\boldsymbol{w},z^{(i)}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-z^{(i)})) \quad \text{s.t.} \quad z^{(i)} = y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})$$

Examples: Logistic Regression

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})))$$

Reformulate:

$$\min_{\boldsymbol{w},z^{(i)}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-z^{(i)})) \quad \text{s.t.} \quad z^{(i)} = y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})$$

Lagrangian:

Examples: Logistic Regression

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})))$$

Reformulate:

$$\min_{\boldsymbol{w},z^{(i)}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-z^{(i)})) \quad \text{s.t.} \quad z^{(i)} = y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})$$

Lagrangian:

$$\begin{aligned} L() &= \frac{C}{2}\|\boldsymbol{w}\|_2^2 - \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)}) \\ &+ \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left[\log(1+\exp(-z^{(i)})) + \lambda^{(i)} z^{(i)}\right] \end{aligned}$$

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$\frac{\partial L}{\partial \boldsymbol{w}}$:

$\frac{\partial L}{\partial z^{(i)}}$:

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$\frac{\partial L}{\partial \boldsymbol{w}}$: $$C\boldsymbol{w} = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})$$

$\frac{\partial L}{\partial z^{(i)}}$:

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$$\frac{\partial L}{\partial \boldsymbol{w}}: \qquad C\boldsymbol{w} = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})$$

$$\frac{\partial L}{\partial z^{(i)}}: \qquad \lambda^{(i)} = \frac{\exp(-z^{(i)})}{1+\exp(-z^{(i)})} \qquad \Longrightarrow \lambda^{(i)} \geq 0$$

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$$\frac{\partial L}{\partial \boldsymbol{w}}: \qquad C\boldsymbol{w} = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})$$

$$\frac{\partial L}{\partial z^{(i)}}: \qquad \lambda^{(i)} = \frac{\exp(-z^{(i)})}{1+\exp(-z^{(i)})} \qquad \Longrightarrow \lambda^{(i)} \geq 0$$

$$\Longrightarrow z^{(i)} = \log\frac{1-\lambda^{(i)}}{\lambda^{(i)}} \qquad \Longrightarrow \lambda^{(i)} \leq 1$$

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$$\frac{\partial L}{\partial \boldsymbol{w}}: \qquad C\boldsymbol{w} = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})$$

$$\frac{\partial L}{\partial z^{(i)}}: \qquad \lambda^{(i)} = \frac{\exp(-z^{(i)})}{1+\exp(-z^{(i)})} \qquad \Longrightarrow \lambda^{(i)} \geq 0$$

$$\Longrightarrow z^{(i)} = \log \frac{1-\lambda^{(i)}}{\lambda^{(i)}} \qquad \Longrightarrow \lambda^{(i)} \leq 1$$

Dual function:

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$$\frac{\partial L}{\partial \boldsymbol{w}}: \qquad C\boldsymbol{w} = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})$$

$$\frac{\partial L}{\partial z^{(i)}}: \qquad \lambda^{(i)} = \frac{\exp(-z^{(i)})}{1+\exp(-z^{(i)})} \qquad \Longrightarrow \lambda^{(i)} \geq 0$$

$$\Longrightarrow z^{(i)} = \log \frac{1-\lambda^{(i)}}{\lambda^{(i)}} \qquad \Longrightarrow \lambda^{(i)} \leq 1$$

Dual function:

$$g(\lambda) = -\frac{1}{2C}\|\sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} H(\lambda^{(i)})$$

with binary entropy $H(\lambda^{(i)})$

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$$\frac{\partial L}{\partial \boldsymbol{w}}: \qquad C\boldsymbol{w} = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})$$

$$\frac{\partial L}{\partial z^{(i)}}: \qquad \lambda^{(i)} = \frac{\exp(-z^{(i)})}{1+\exp(-z^{(i)})} \qquad \Longrightarrow \lambda^{(i)} \geq 0$$

$$\Longrightarrow z^{(i)} = \log \frac{1-\lambda^{(i)}}{\lambda^{(i)}} \qquad \Longrightarrow \lambda^{(i)} \leq 1$$

Dual function:

$$g(\lambda) = -\frac{1}{2C}\| \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} H(\lambda^{(i)})$$

with binary entropy $H(\lambda^{(i)})$

Dual program:

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$$\frac{\partial L}{\partial \boldsymbol{w}}: \qquad C\boldsymbol{w} = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})$$

$$\frac{\partial L}{\partial z^{(i)}}: \qquad \lambda^{(i)} = \frac{\exp(-z^{(i)})}{1+\exp(-z^{(i)})} \qquad \Longrightarrow \lambda^{(i)} \geq 0$$

$$\Longrightarrow z^{(i)} = \log \frac{1-\lambda^{(i)}}{\lambda^{(i)}} \qquad \Longrightarrow \lambda^{(i)} \leq 1$$

Dual function:

$$g(\lambda) = -\frac{1}{2C}\|\sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} H(\lambda^{(i)})$$

with binary entropy $H(\lambda^{(i)})$
Dual program:

$$\max_{\lambda} g(\lambda) \quad \text{s.t.} \quad 0 \leq \lambda^{(i)} \leq 1 \qquad \forall (x^{(i)}, y^{(i)}) \in \mathcal{D}$$

Instead of optimizing the primal we can optimize the dual and convert the result

Instead of optimizing the primal we can optimize the dual and convert the result

Why is this useful?

Instead of optimizing the primal we can optimize the dual and convert the result

Why is this useful?

- Sometimes less constraints

Instead of optimizing the primal we can optimize the dual and convert the result

Why is this useful?

- Sometimes less constraints
- Sometimes easier to optimize

Instead of optimizing the primal we can optimize the dual and convert the result

Why is this useful?

- Sometimes less constraints
- Sometimes easier to optimize
- Sometimes interesting insights

Instead of optimizing the primal we can optimize the dual and convert the result

Why is this useful?

- Sometimes less constraints
- Sometimes easier to optimize
- Sometimes interesting insights
- Sometimes lower bounds

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda,\nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda,\nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

- May only have simple constraints if at all

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda,\nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

- May only have simple constraints if at all
- Can be used for sensitivity analysis

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda, \nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

- May only have simple constraints if at all
- Can be used for sensitivity analysis
- Lower-bounds the optimal primal value

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda, \nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

- May only have simple constraints if at all
- Can be used for sensitivity analysis
- Lower-bounds the optimal primal value
- Dual Program is always concave:

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda,\nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

- May only have simple constraints if at all
- Can be used for sensitivity analysis
- Lower-bounds the optimal primal value
- Dual Program is always concave:

$$g(\lambda,\nu) = \min_{\boldsymbol{w}\in\mathcal{W}} L(\boldsymbol{w},\lambda,\nu) := f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda, \nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

- May only have simple constraints if at all
- Can be used for sensitivity analysis
- Lower-bounds the optimal primal value
- Dual Program is always concave:

$$g(\lambda, \nu) = \min_{\boldsymbol{w} \in \mathcal{W}} L(\boldsymbol{w}, \lambda, \nu) := f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

  - Pointwise minimum

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda,\nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

- May only have simple constraints if at all
- Can be used for sensitivity analysis
- Lower-bounds the optimal primal value
- Dual Program is always concave:

$$g(\lambda,\nu) = \min_{\boldsymbol{w}\in\mathcal{W}} L(\boldsymbol{w},\lambda,\nu) := f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

  - Pointwise minimum
  - Affine functions in λ,ν

Weak duality:

$$f(\boldsymbol{w}^*) \geq g(\lambda^*, \nu^*)$$

Weak duality:

$$f(\boldsymbol{w}^*) \geq g(\lambda^*, \nu^*)$$

- Always holds (for convex and non-convex problems)

Weak duality:

$$f(\boldsymbol{w}^*) \geq g(\lambda^*, \nu^*)$$

- Always holds (for convex and non-convex problems)
- Can be used to find nontrivial lower bounds

Weak duality:

$$f(\boldsymbol{w}^*) \geq g(\lambda^*, \nu^*)$$

- Always holds (for convex and non-convex problems)
- Can be used to find nontrivial lower bounds

Strong duality:

$$f(\boldsymbol{w}^*) = g(\lambda^*, \nu^*)$$

Weak duality:

$$f(\boldsymbol{w}^*) \geq g(\lambda^*, \nu^*)$$

- Always holds (for convex and non-convex problems)
- Can be used to find nontrivial lower bounds

Strong duality:

$$f(\boldsymbol{w}^*) = g(\lambda^*, \nu^*)$$

- Does not hold in general

Weak duality:

$$f(\boldsymbol{w}^*) \geq g(\lambda^*, \nu^*)$$

- Always holds (for convex and non-convex problems)
- Can be used to find nontrivial lower bounds

Strong duality:

$$f(\boldsymbol{w}^*) = g(\lambda^*, \nu^*)$$

- Does not hold in general
- (Usually) holds for convex problems

Quiz:

Quiz:

- What to do before computing the Lagrangian?

Quiz:

- What to do before computing the Lagrangian?
- How to obtain the dual program?

Quiz:

- What to do before computing the Lagrangian?
- How to obtain the dual program?
- Why duality?

Important topics of this lecture

Important topics of this lecture

- Lagrangian

Important topics of this lecture

- Lagrangian
- Dual program

Important topics of this lecture

- Lagrangian
- Dual program

Up next:

- Support vector machines