

Numerical Optimisation  
Nonsmooth optimisation

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat [edu\\_assist\\_pro](#)

Department of Computer Science  
Centre for Medical Image Computing  
Centre for Inverse Problems  
University College London

Lecture 16

# Subgradient

For convex differentiable function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  it holds

$$f(y) \geq f(x) + \nabla f(x)^T(y - x).$$

Assignment Project Exam Help

A vector  $g \in \mathbb{R}^n$  is a **subgradient** of a function  $f$  at  $x \in \text{dom } f$  if

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom } f.$$

- <https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pr

Figure:  $\partial f(x_1) = \{\nabla f(x_1)\} = \{g_1\}$ ,  $\partial f(x_2) = [g_3, g_2]$ . Fig. from S. Boyd, EE364b, Stanford University.

A function  $f$  is called **subdifferentiable** at  $x$  if there exists at least one subgradient at  $x$ .

**Subdifferential** of  $f$  at  $x$ ,  $\partial f(x)$ , is the set of all subgradients of  $f$  at  $x$ .

$\partial f(x)$  is nonempty if  $f$  is convex.

**Properties**  
• If  $f$  is convex, then  $\partial f(x)$  is nonempty for  $x \in \text{relint dom } f$ .

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro

If  $f(x)$  is convex

- $\partial f(x)$  is nonempty for  $x \in \text{relint dom } f$
- then  $f$  is continuous at  $x$ , and hence the  $\partial f(x)$  is bounded
- $\partial f(x) = \{\nabla f(x)\}$  iff  $f$  differentiable at  $x$

# Minimum of nondifferentiable function (unconstraint)

A point  $x^*$  is a minimiser of a function  $f$  (not necessarily convex)

iff  $f$  is subdifferentiable at  $x^*$  and

$$0 \in \partial f(x^*),$$

i.e.  $g$

**Pro**

$f$  is subdifferentiable at  $x^*$  with  $0 \in \partial f$

$f(x) \geq f(x^*) + 0^T(x - x^*)$  for all  $x \in$

The condition  $0 \in \partial f(x^*)$  reduces to

and differentiable at  $x^*$ . Note, that in that case also it is a necessary and sufficient condition.

# Minimum of nondifferentiable function (constraint)

Convex constraint optimisation problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad (\text{COP})$$

subject to  $f_i(x) \leq 0, \quad i = 1, \dots, m,$

where

•

•

**Gen**

$x^*$  is primal optimal and  $\lambda^*$  dual optima

$$\begin{aligned} f(x^*) &\leq 0, \\ \lambda_i^* &\geq 0, \end{aligned} \quad (\text{KT})$$

$$0 \in \partial f(x^*) + \sum_{i=1}^m \lambda_i^* \partial f_i(x^*),$$

$$\lambda_i^* f_i(x^*) = 0$$

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro

# Directional derivatives and subdifferential

For a convex function the *directional derivative* at  $x$  in the direction  $v$  is

$$f'(x; v) = \lim_{t \rightarrow 0} \frac{f(x + tv) - f(x)}{t}$$

The limit always exists for a convex function, though it can be  $\pm\infty$ . I

$f$  is convex  
 $v \in \mathbb{R}^n$

The directional derivative  $f'(x; v)$  of a convex

$f'(x; v) = \sup_{g \in \partial f(x)} g^T v$

**Proof idea:** Note that  $f'(x; v) \geq \sup_{g \in \partial f(x)} g^T v$  by the definition of the subgradient  $f(x + tv) - f(x) \geq tg^T v$  for any  $t \in \mathbb{R}$  and  $g \in \partial f(x)$ . Other direction: show that all affine functions below  $v \rightarrow f'(x; v)$  may be taken to be linear.

**Weak subgradient calculus:** formulas for finding *one*  $g \in \partial f(x)$ .

If you can compute  $f$ , you can usually compute one subgradient.

Many algorithms require only one subgradient.

**Strong subgradient calculus:** formula for finding *the whole* subdifferential  $\partial f(x)$

Opti

diffe

Basic rules:

- scaling: for  $\alpha \geq 0$ ,  $\partial(\alpha f) = \alpha \partial f$
- addition:  $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- affine transformation:  $g(x) = f(Ax + b)$   
 $\partial g(x) = A^T \partial f(Ax + b)$
- finite point wise maximum:  $f = \max_{i=1, \dots, m} f_i$ ,  
 $\partial f(x) = \text{Co} \bigcup \{ \partial f_i(x) : f_i(x) = f(x) \}$  (convex hull of a union of subdifferentials of active functions at  $x$ )

## Subgradient and descent direction

$p$  is a descent direction for  $f$  at  $x$  if  $f'(x; p) < 0$ .

If  $f$  is differentiable,  $-\nabla f$  is always a descent direction (except when it is 0).

For a nondifferentiable convex function  $f$ ,  $p = g, g \in \partial f(x)$  need not to be

Exa

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro

Figure: Fig. from S. Boyd, EE364b, Stanford University.



For a convex  $f$ , if  $f(z) < f(x)$ ,  $g \in \partial f(x)$ , then for small  $t > 0$

**Assignment Project Exam Help**

Thus  $g$  is descent direction for  $x - z$ , for any  $z$  with  $f(z) < f(x)$

**Pro** <https://eduassistpro.github.io>

$$\|x - tg - z\|_2^2 = \|x - z\|_2^2 - 2tg^T(x - z)$$

**Add WeChat** [edu\\_assist\\_pro](https://eduassistpro.github.io)

$$t: \frac{1}{2}\|g\|_2^2 < f(x) - f(z)$$

In particular, choosing  $z = x^*$ , we obtain that the negative subgradient is a descent direction for distance to optimal point  $x^*$ .

**Proximal operator** of  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$

$$\text{prox}_{\lambda f}(v) := \arg \min_x (f(x) + 1/(2\lambda) \|x - v\|_2^2), \lambda > 0 \quad (\text{PROX})$$

Evaluating  $\text{prox}_{\lambda f}$  involves solving a convex optimisation problem

# Assignment Project Exam Help

Can e

prob

linea

<https://eduassistpro.github.io>

Indicator function of a closed convex set,  $I_C$

# Add WeChat edu\_assist\_pro

$$I_C(v) = \begin{cases} 0 & \text{if } v \in C \\ \infty & \text{otherwise} \end{cases}$$

Proximal operator of  $I_C$  is the Euclidean projection

$$\text{prox}_{\lambda I_C}(v) = \arg \min_{x \in C} \|x - v\|_2 = \Pi_C(v)$$

Many properties of projection carry over to proximal operator.

# Examples of proximal operators

Important special choices of  $f$ , for which  $\text{prox}_{\lambda f}$  has a closed form:

- $f(x) = \frac{1}{2} \|Px - q\|_2^2$ ,

$$\text{prox}_{\lambda f}(v) = (P^T P + \lambda^{-1} I)^{-1} (P^T q + \lambda^{-1} v).$$
$$(P^T P + Q)^{-1} = Q^{-1} - Q^{-1} P^T (I + P Q^{-1} P^T)^{-1} P Q^{-1}.$$

- 

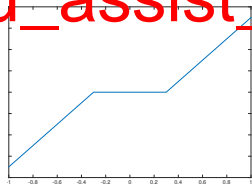
<https://eduassistpro.github.io>

- $f(x) = \|x\|_1$

$\text{prox}_{\lambda}(v) = S_{\lambda}(v)$

with elementwise soft thresholding

$$S_{\delta}(x) = \begin{cases} x - \delta & x > \delta \\ 0 & x \in [-\delta, \delta] \\ x + \delta & x < -\delta \end{cases}$$



# Examples of proximal operators

Another important example which does not admit close form is  
Total Variation,  $f(x) = TV(x)$ , defined as follows

$$TV(x) := \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \sqrt{(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2}$$

<https://eduassistpro.github.io>

assuming standard reflexive boundary conditions

$$x_{m+1,j} = x_{m,j}, \quad x_{i,n+1} = x_{i,n}$$

The proximal operator has to be computed iteratively using  
e.g. Chambolle-Pock algorithm (primal dual proximal gradient).

Proximal operator

$$\operatorname{prox}_{\lambda f}(v) = \arg \min_x (f(x) + 1/(2\lambda) \|x - v\|_2^2).$$

Assignment Project Exam Help

The first order condition for the minimiser reads

<https://eduassistpro.github.io>

$$v - x \in \lambda \partial f(x)$$

$$\operatorname{prox}_{\lambda f}(v) = (I + \lambda \partial f)^{-1}(v)$$

Mapping  $(I + \lambda \partial f)^{-1}$  is called **resolve**

Add WeChat edu\_assist\_pro

$x^*$  minimises  $f$  iff  $x^*$  is a fixed point

$$x^* = \operatorname{prox}_f(x^*)$$

## Moreau envelope or Moreau-Yosida regularisation of $f$

$$M_{\lambda f}(v) = \inf_x \left( f(x) + 1/(2\lambda) \|x - v\|_2^2 \right).$$

Assignment Project Exam Help

$M_{\lambda f}$  is a smoothed (regularised) version of  $f$ .

- 

- 

- 

<https://eduassistpro.github.io>

Can show that  $M_f = (f^* + 1/2 \|\cdot\|_2^2)^*$ .

Example: Moreau envelope of  $|\cdot|$  is the

Add WeChat edu\_assist\_pro

$$M_{|\cdot|}(x) = \begin{cases} x^2 & |x| \leq 1 \\ 2|x| - 1 & |x| > 1 \end{cases}$$

**Moreau decomposition:**  $v = \text{prox}_f(v) + \text{prox}_{f^*}(v)$  is  
generalisation of orthogonal decomposition  $v = \Pi_W(v) + \Pi_{W^\perp}(v)$ .  
It follows from Moreau decomposition that  $(I_W)^* = I_{W^\perp}$ .

$$\min_x f(x) + g(x) \quad \text{subject to } x \in \mathbb{E} \quad (1)$$

•  $\mathbb{E}$ : finite dimensional Euclidean space with inner product  $\langle \cdot, \cdot \rangle$  and self dual norm  $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2} = \|\cdot\|_*$ ,

• <https://eduassistpro.github.io>

$$\|\nabla f(x) - \nabla f(y)\| \leq L(f)\|x - y\|, \quad x, y \in \mathbb{E}.$$

•  $g : \mathbb{E} \rightarrow (-\infty, \infty]$  proper closed convex

From first order optimality condition we have

$$\begin{aligned} 0 &\in \nabla f(x^*) + \\ 0 &\in \tau \nabla f(x^*) + \tau \partial g(x^*) - x^* + x^* \\ (I + \tau \partial g)(x^*) &\in (I - \tau \nabla f)(x^*) \\ x^* &= (I + \tau \partial g)^{-1}(I - \tau \nabla f)(x^*) \end{aligned} \quad (2)$$

$$x_k = \text{prox}_{\tau_k g}(x_{k-1} - \tau_k \nabla f(x_{k-1}))$$

Assignment Project Exam Help



<https://eduassistpro.github.io>

- **Proximal Minimization:**  $f(x) = 0$ : n  
minimisation

Add WeChat edu\_assist\_pro

$$x_k = \arg \min_x \left\{ g(x) + \frac{1}{2\tau_k} \|x_{k-1} - \tau_k \nabla f(x_{k-1})\|^2 \right\}.$$

- **Iterative Shrinkage Thresholding Algorithm (ISTA):**  
 $g(x) = \|x\|_1, f(x) = \|Ax - b\|^2, \tau_k \in (0, 2/L(f))$

$$x_k = S_{\tau_k}(x_{k-1} - \tau_k \nabla f(x_{k-1})).$$



$$x_k = \text{prox}_{\tau_k g}(x_{k-1} - \tau_k \nabla f(x_{k-1}))$$

Assignment Project Exam Help



<https://eduassistpro.github.io>

- **Iterative Shrinkage Thresholding Algorithm**

$g(x) = \|\lambda\|_1, f(x) = \|Ax - b\|_2^2$

$$x_k = S_{\tau_k}(x_{k-1} - \tau_k \nabla f(x_{k-1}))$$

Slow convergence, if  $\tau_k = \tau = 1/L, L \geq L(f)$

$$F(x_k) - F^* \leq \frac{L \|x_0 - x^*\|^2}{2k}.$$

# Fast Iterative Shrinkage Thresholding Algorithm (FISTA):

Initialize:  $y_1 := x_0 \in \mathbb{E}$ ,  $\tau_1 = 1$ .

$$\text{Step } k : \quad x_k = \text{prox}_{1/L}(g) \left( y_k - \frac{1}{L} \nabla f(y_k) \right)$$

Assignment Project Exam Help

$$\tau_{k+1} = \frac{1 + 4\tau_k^2}{2}$$

Conv

<https://eduassistpro.github.io>

Add WeChat [edu\\_assist\\_pro](#)

Fast Projection Gradient [Nesterov'83]:

$$x_k = \Pi_{\mathcal{C}} \left( y_k - \frac{1}{L} \nabla f(y_k) \right).$$

More details on Nesterov algorithm see e.g. <http://www.seas.ucla.edu/~vandenbe/236C/lectures/fgrad.pdf>

# Review: Optimisation with equality constraints

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ , closed, proper and convex.

Primal problem

Assignment Project Exam Help

Lagrangian

Dual problem

<https://eduassistpro.github.io>

$$g(y) = \inf_x \mathcal{L}(x, y) = -f^* (A^T y) - b^T y$$

$y$ : dual variable (Lagrange multiplier),

$f^*$ : convex conjugate of  $f$  ( $f^*$  is convex not).

Add WeChat: edu\_assist\_pro

Dual problem (always concave,  $y^* \leq x^*$ ,  $y^* = x^*$  if strong duality holds)

$$\max_y g(y). \quad (3)$$

**Gradient descent** for primal problem (assuming  $f$  continuously differentiable)

**Assignment Project Exam Help**

**Gradient ascent** for dual problem (assuming  $g$  continuously differentiable)

<https://eduassistpro.github.io>

**Add WeChat edu\_assist\_pro**

Remark: the primal update is part of evaluation of

- + for separable  $f$  it leads to a parallel algorithm.
- various conditions necessary for convergence e.g. strict convexity of  $f$ ,  $f(x) < \infty, \forall x$ .

# Augmented Lagrangian

## Augmented Lagrangian

$$\mathcal{L}_\rho(x, y) = f(x) + y^T(Ax - b) + \rho/2 \|Ax - b\|_2^2, \quad \rho > 0 \quad (\text{AL})$$

Equivalent to Lagrangian of an equivalent problem (for all feasible  $x$  the quadratic term equals 0)

2

Met

<https://eduassistpro.github.io>

$k+1$

$\rho$

$k$

$$y_{k+1} = y_k + \rho(A$$

Add WeChat [edu\\_assist\\_pro](#)

Using  $\rho$  at a step size guarantees dual feasibility  $k+1 \quad k+1$ :

$$0 = \nabla_x \mathcal{L}_\rho(x_{k+1}, y_k) = \nabla f(x) + A^T y_k + \rho A^T(Ax - b) \Big|_{x=x_{k+1}} =$$

$$\nabla f(x_{k+1}) + A^T y_{k+1} =: s_{k+1} = 0.$$

+ converges under more general conditions

– augmented Lagrangian is non-separable.

# Alternating Directions Methods of Multipliers (ADMM)

Blend separability of dual ascent with superior convergence of MM:

$$\min_{x \in \mathbb{R}^n, z \in \mathbb{R}^m} f(x) + g(z) \quad \text{subject to } Ax + Bz = c \quad (4)$$

Assignment Project Exam Help  
with  $A \in \mathbb{R}^{p \times n}$ ,  $B \in \mathbb{R}^{p \times m}$ ,  $c \in \mathbb{R}^p$  and  $g: \mathbb{R}^m \rightarrow \mathbb{R} \cup \{\infty\}$  closed, proper and convex.

The e

and

Aug

$$L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2} \|Ax + Bz - c\|_2^2,$$

ADMM

Add WeChat edu\_assist\_pr

$$x_{k+1} = \arg \min_x L_\rho(x, z_k, y_k)$$

$$z_{k+1} = \arg \min_z L_\rho(x_{k+1}, z, y_k)$$

$$y_{k+1} = y_k + \rho(Ax_{k+1} + Bz_{k+1} - c).$$

# Alternating Directions Methods of Multipliers (ADMM)

Blend separability of dual ascent with superior convergence of MM:

$$\min_{x \in \mathbb{R}^n, z \in \mathbb{R}^m} f(x) + g(z) \quad \text{subject to } Ax + Bz = c \quad (4)$$

with  $A \in \mathbb{R}^{p \times n}$ ,  $B \in \mathbb{R}^{p \times m}$ ,  $c \in \mathbb{R}^p$  and  $g: \mathbb{R}^m \rightarrow \mathbb{R} \cup \{+\infty\}$  closed, proper and convex.

The e  
and

Augmented Lagrangian

$$L_\rho(x, z, y) = f(x) + g(z) + y^T (Ax + Bz - c) + \frac{\rho}{2} \|Ax + Bz - c\|_2^2$$

**Dual ascent** on  $\mathcal{L}_\rho$  (joint minimisation)

$$(x_{k+1}, z_{k+1}) = \arg \min_{x, z} L_\rho(x, z, y_k)$$

$$y_{k+1} = y_k + \rho(Ax_{k+1} + Bz_{k+1} - c).$$

## ADMM: scaled form

Augmented Lagrangian

$$\begin{aligned}
 L_\rho(x, z, y) &= f(x) + g(z) + y^T (Ax + Bz - c) + \rho/2 \|Ax + Bz - c\|_2^2 \\
 &= f(x) + g(z) + y^T r + \rho/2 \|r\|_2^2.
 \end{aligned}$$

Assignment Project Exam Help

$\|_2^2$ ,

<https://eduassistpro.github.io>

with

ADMM: scaled form

Add WeChat edu\_assist\_pro

$$x_{k+1} = \arg \min_x f(x) + \rho/2 \|Ax_{k+1} + Bz_k - c + u_k\|_2^2$$

$$z_{k+1} = \arg \min_z g(z) + \rho/2 \|Ax_{k+1} + Bz - c + u_k\|_2^2$$

$$u_{k+1} = u_k + Ax_{k+1} + Bz_{k+1} - c.$$



Assume in addition that the unaugmented Lagrangian  $\mathcal{L}$  has a saddle point.

Assignment Project Exam Help

For  $f, g$  proper, closed, convex, it follows that strong duality holds (no explicit assumptions on  $A, B, c$ ).

Under

- <https://eduassistpro.github.io> approach feasibility.
- Objective convergence:  $f(x^k) + g^*(z^k)$  approaches the objective function of the iterates approach
- Dual variable convergence:  $y^k \rightarrow y^*$  is a dual optimal point.

Note, that  $x^k, z^k$  need not converge to optimal points, although such a result can be shown under additional assumptions.

# Optimality conditions

Necessary and sufficient optimality conditions for ADMM

$$Ax^* + Bz^* - c = 0 \quad \text{primal feasibility}$$

$$0 \in \partial f(x^*) + A^T y^* \quad \text{dual feasibility}$$

$$0 \in \partial g(z^*) + B^T y^* \quad \text{dual feasibility}$$

As for

) that

$z_{k+1}$

From

$$0 \in \partial f(x_{k+1}) + A^T y_k + \rho A^T (Ax_{k+1} + Bz_{k+1} - c)$$

$$= \partial f(x_{k+1}) + A^T (y_k + \rho r_{k+1})$$

$$= \partial f(x_{k+1}) + A^T y_{k+1} + \rho A^T (Ax_{k+1} + Bz_{k+1} - c)$$

or equivalently

$$s_{k+1} := \rho A^T B(z_{k+1} - z_k) \in \partial f(x_{k+1}) + A^T y_{k+1},$$

which can be interpreted as dual feasibility condition and  $s_{k+1}$  is the *dual residual* at iteration  $k + 1$ .

- S. Boyd, Stanford EE364b  
<http://stanford.edu/class/ee364b/lectures.html>
- L. Vandenbergh, UCLA EE236C  
<http://www.seas.ucla.edu/~vandenbe/ee236c.html>

<https://eduassistpro.github.io>

- A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, A. Beck, M. Teboulle, 2
- Fast Gradient-Based Algorithms for Co-Variation Image Denoising and Deblurring, M. Teboulle, 2009
- A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging, A. Chambolle, T. Pock, 2011