Numerical Optimisation
Nonsmooth optimisation

Department of Computer Sc
Centre for Medical Image Com
Centre for Inverse Proble
University College London

Lecture 16

Subgradient

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

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

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

$$f(y) \quad f(x) + g^{\mathrm{T}}(y \quad x) \quad\quad y \quad \operatorname{dom} f.$$

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.

Subdifferential

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.

$\partial f(x$ nvex.

Pro

halfs

$$\partial f(x) = \bigcap_{z \in \mathrm{dom}\, f} \{g : f(z) \geq$$

If $f(x)$ is convex

- $\partial f(x)$ is nonempty for $x \in \mathrm{relint\,dom}\, f$
- 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 $x^\star$ is a minimiser of a function f (not necessarily convex) iff f is subdifferentiable at $x^\star$ and

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

i.e. g

Pro dom f.

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

$f(x) \geq f(x^\star) + 0^{\mathrm{T}}(x - x^\star)$ for all $x \in$

The condition $0 \in \partial f(x^\star)$ reduces to ex and differentiable at $x^\star$. Note, that in that case also it is a necessary and sufficient condition.

Convex constraint optimisation problem

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

$$\text{subject to} \quad f_i(x) \le 0, \quad i = 1,\dots,m,$$

where

-
-

Gen

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

$$f_i(x^\star) \le 0, \qquad \text{(KKT)}$$

$$\lambda_i^\star \ge 0,$$

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

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

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

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

The limit always exists for a convex function, thought it can be $\pm\infty$. I [illegible] sts.

f is d [illegible] $v \in$ [illegible] $v)$.

The directional derivative $f'(x;v)$ of a con [illegible] es

$$f'(x;v) = \sup_{g \in \partial f(x} \ldots$$

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

Opti

diffe

Basic rules:

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

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, $p = \;\; g$, $g \quad \partial f(x)$ need not to b

Exa

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

Subgradient and distance to sublevel set

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

$$\|x - tg - z\|_2 < \|x - z\|_2.$$

Thus $-g$ is descent direction for $\|x - z\|_2$, for any z with $f(z) < f(x)$

Pro

$$\|x - tg - z\|_2^2 = \|x - z\|_2^2 - 2tg^{\mathrm{T}}(x$$

$$\le \|x - z\|_2^2 - 2t(f(x)$$

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

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

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

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

Evaluating prox_f involves solving a convex optimisation problem.

Can e
prob
linea

Indicator function of a closed convex set, /

$$I_{\mathcal{C}}(x) = \begin{cases} 0 & \\ \infty & \text{ot} \end{cases}$$

Proximal operator of $I_{\mathcal{C}}$ is the Euclidean projection

$$\operatorname{prox}_{\lambda I_{\mathcal{C}}}(v) = \arg\min_{x \in \mathcal{C}} \|x - v\|_2 = \Pi_{\mathcal{C}}(v)$$

Many properties of projection carry over to proximal operator.

Examples of proximal operators

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

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

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

$$(P^{\mathrm{T}}P + Q)^{-1} = Q^{-1} - Q^{-1}P^{\mathrm{T}}(I + PQ^{-1}P^{\mathrm{T}})^{-1}PQ^{-1}.$$

-

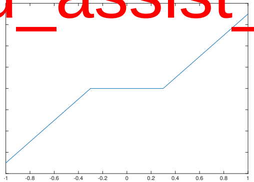
i

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

$$\mathrm{prox}_{\lambda f}(v) = S_{\lambda}(v),$$

with elementwise soft thresholding

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



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

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

$$\sum_{i=1} \cdots \sum_{i=1} \cdots 1|$$

assuming standard reflexive boundary conditi

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

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

Proximal operator

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

The first order condition for the minimiser reads

$$\begin{aligned} v - x &\in \lambda \partial f(x) \\ \text{prox}_{\lambda f}(v) &= (I + \lambda \end{aligned}$$

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

$x^\star$ minimises f iff $x^\star$ is a fixed point

$$x^\star = \text{prox}_f(x^\star)$$

Moreau envelope or **Moreau-Yosida regularisation** of f

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

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

-
-
-

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

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

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

Moreau decomposition: $v = \mathrm{prox}_f(v) + \mathrm{prox}_{f^\star}(v)$ is generalisation of orthogonal decomposition $v = \Pi_W(v) + \Pi_{W^\perp}(v)$.

It follows from Moreau decomposition that $(I_W)^\star = I_{W^\perp}$.

Forward Backward splitting

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

- $\mathbb{E}$: finite dimensional Euclidean space with inner product $< \cdot, \cdot >$ and self dual norm $\| \cdot \| =< \cdot, \cdot >^{1/2} = \| \cdot \|_*$,
-
  $\|\nabla f(x) - \nabla f(y)\| \leq L(f)\|x - y\|, \quad x, y$.
- $g : \mathbb{E} \to (-\infty, \infty]$ proper closed conve

From first order optimality condition we have

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

$$\begin{aligned} x_k &= \operatorname{prox}_{\tau_k g}(x_{k-1} - \tau_k \nabla f(x_{k-1})) \\ &= \arg\min_x \left\{ g(x) + \frac{1}{2\tau_k} \|x - (x_{k-1} - \tau_k \nabla f(x_{k-1}))\|^2 \right\}. \end{aligned}$$

-

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

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

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

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

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

-

  $$\quad_{k} \quad_{\mathcal{C}} \quad_{k-1} \quad_{k} \quad_{k-1}$$

- **Iterative Shrinkage Thresholding Al**

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

  $$x_k = S_{\tau_k}(x_{k-1} - \tau_k \quad\quad _{k-1}$$

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

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

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

Step k :

$$x_k = \mathrm{prox}_{1/L}(g)\left(y_k - \frac{1}{L}\nabla f(y_k)\right)$$

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

Conv

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

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`

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

Primal problem

$$\min_x f(x) \quad \text{subject to } Ax = b$$

Lagrangian

Dual

$$g(y) = \inf_x \mathcal{L}(x, y) = -f \;(\;\; A \;\; y) \quad b \;\; y$$

y: dual variable (Lagrange multiplier),
f^*: convex conjugate of f (f^* is conve
not).

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

$$\max_y g(y). \tag{3}$$

Gradient descent for primal problem (assuming f continuously differentiable)

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

Gradient ascent for dual problem (assuming g continuously diffe

$k+1$ k k $k+1$

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

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

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

2

Met

$k+1$ ρ k

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

Using ρ at a step size guarantees dual feasibility $\quad {}_{k+1} \quad {}_{k+1})$:

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

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

\+ converges under more general conditions

– augmented Lagrangian is non-separable.

Blend separability of dual ascent with superior convergence of MM:

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

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

The e x

and

Aug

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

ADMM

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

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

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

Blend separability of dual ascent with superior convergence of MM:

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

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

The e ... x
and

Augmented Lagrangian

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

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

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

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

Augmented Lagrangian

$$L_\rho(x,z,y) = f(x) + g(z) + y^{\mathrm{T}} \underbrace{(Ax + Bz - c)}_{=:r} + \rho/2 \| \underbrace{Ax + Bz - c}_{} \|_2^2,$$

$$= f(x) + g(z) + y^{\mathrm{T}} r + \rho/2 \|r\|_2^2.$$

$$\|_2^2,$$

with

ADMM: scaled form

$$x_{k+1} = \arg\min_x f(x) + \rho/2 \|Ax \quad {}_2$$

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

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

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

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

Und

- approach feasibility.
- Objective convergence: $f(x^k) + g$ he objective function of the iterates approac

- Dual variable convergence: $y^k \to$ 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.

Necessary and sufficient optimality conditions for ADMM

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

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

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

As fo) that

z_{k+1}

Fro

$$0 \in \partial f(x_{k+1}) + A^{\mathrm{T}} y_k + \rho A^{\mathrm{T}}(Ax \quad + Bz \quad c)$$

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

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

or equivalently

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

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. Vandenberghe, UCLA EE236C
  http://www.seas.ucla.edu/~vandenbe/ee236c.html
-
- 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 Deblurri
  M. Teboulle, 2009
- A First-Order Primal-Dual Algorithm for Convex Problems withApplications to Imaging, A. Chambolle, T. Pock, 2011