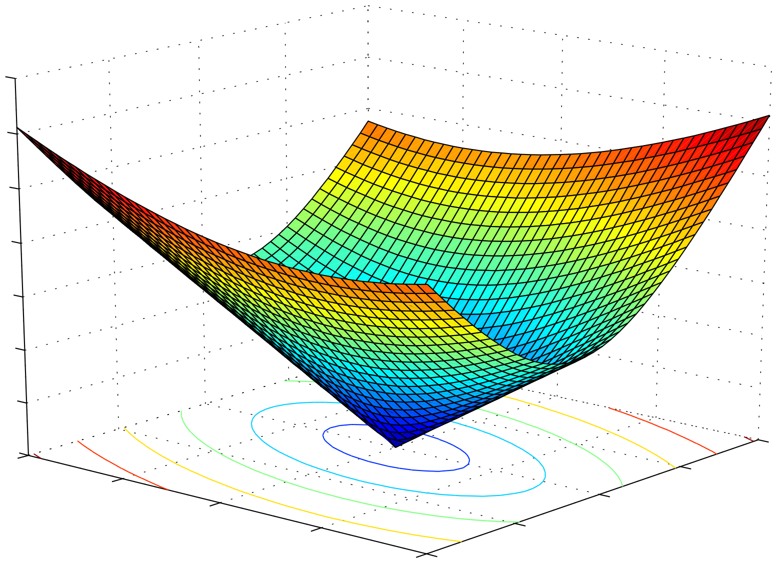
Variational Methods II

Convex Relaxation & Non-smooth Optimization



Dr. Martin Oswald

Computer Vision and Geometry Group

ETH zürich

Discrete vs. Continuous

Energy $E : \mathcal{V} \to \mathbb{R}$ assigns each element of the space $\mathcal{V}$ a real number (energy).

	Vector space $\mathcal{V} = \mathbb{R}^n$	Function space $\mathcal{V} = \mathcal{L}^2(\Omega)$	
Elements	finitely many Elements $x_i, \quad i \in \{1, \dots, n\}$	infinitely many Elements $u(x), \quad x \in \Omega$	
Inner Product	$\langle x, y \rangle = \sum_{i=1}^n x_i y_i$	$\langle u, v \rangle = \int_\Omega uv \; dx$	
Norm	$\lvert x \rvert_2 = \sqrt{\sum_{i=1}^n x_i^2}$	$\lVert u \rVert_2 = \left(\int_\Omega \lvert u \rvert^2 dx \right)^{\frac{1}{2}}$	
Gradient	$dE(x)/dx = \nabla E(x)$	$dE(u)/du = ?$	(Fréchet)
Directional	$\delta E(x; h) = \nabla E(x) \cdot h$	$\delta E(u; h) = ?$	(Gâteaux)
Extrema Condition	$dE(x)/dx = 0$	$dE(u)/du = ?$	

Variational Principle

Theorem

If $\hat{u} \in \mathcal{V}$ is an extremum of a functional $E : \mathcal{V} \to \mathbb{R}$, then

$$\delta E(\hat{u}; h) = 0 \quad \text{for all } h \in \mathcal{V}.$$

- The variational principle is a generalization of the necessary condition for extrema for functions in $\mathbb{R}^n$ to infinite dimensional spaces.
- The necessary conditions must hold for all directions h. Similar to the finite dimensional case, the directional derivative corresponds to the projection of the functional gradient onto the direction h.
- Therefore the Gâteaux derivative can also be written as

$$\delta E(\hat{u}; h) = \left\langle \frac{dE(u)}{du}, h \right\rangle = \int \frac{dE(u)}{du}(x) h(x) \, dx$$

Euler-Lagrange Equation

Theorem

Let $\hat{u} \in \mathcal{V}$ be an extremum of the functional $E : \mathcal{C}^1(\Omega) \to \mathbb{R}$ of the form

$$E(u) = \int_\Omega L(u, \nabla u, x) dx$$

and $L : \Omega \times \Omega^n \times \Omega \to \mathbb{R}, (a, b, x) \mapsto L(a, b, x)$ be continuously differentiable.
Then $\hat{u}$ satisfies the Euler-Lagrange equation

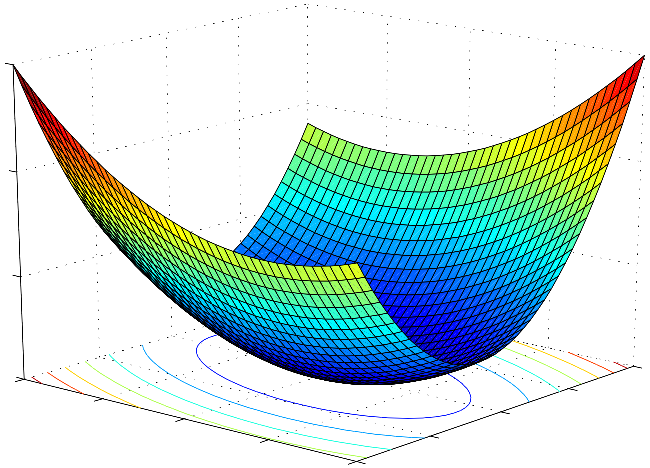
$$\partial_a L(u, \nabla u, x) - \mathrm{div}_x(\nabla_b L(u, \nabla u, x)) = 0$$

where the divergence is computed with respect to the location variable x and

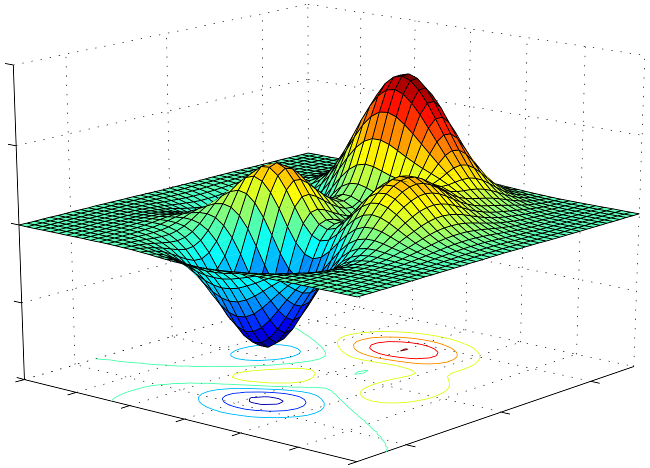
$$\partial_a L := \frac{\partial L}{\partial a} \qquad \nabla_b L := \left[\frac{\partial L}{\partial b_1} \cdots \frac{\partial L}{\partial b_n}\right]^T$$

- The Euler-Lagrange equation is a PDE which has to be satisfied for an extremal point $\hat{u}$.

Why care about Convexity?



convex



non-convex

"In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

Ralph Tyrrell Rockafellar

*(Mathematician and pioneering expert in convex analysis and optimization theory, *1935)*

Convexity

Definition (Convex Set)

A set C is called convex if $\lambda x + (1-\lambda) y \in C$ for all $x, y \in C$ and $0 \leq \lambda \leq 1$.

Geometrically, a set is convex iff the connecting line between any two points in the set is also entirely contained in the set.

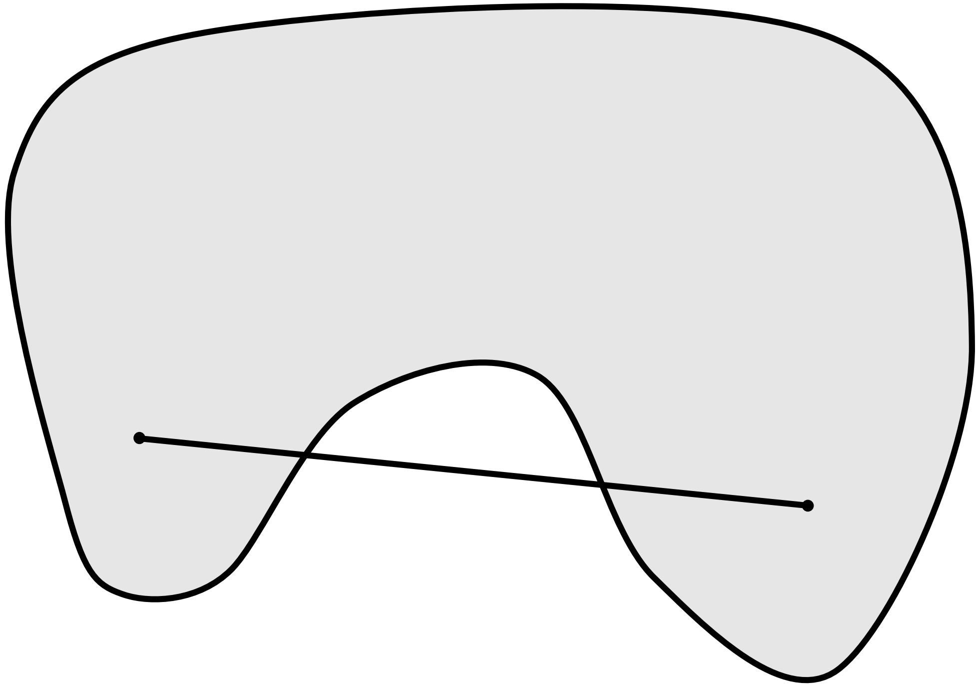
Definition (Epigraph)

The epigraph of a function $f : \mathcal{V} \to \mathbb{R}$ is the set of all point which lie above the graph:

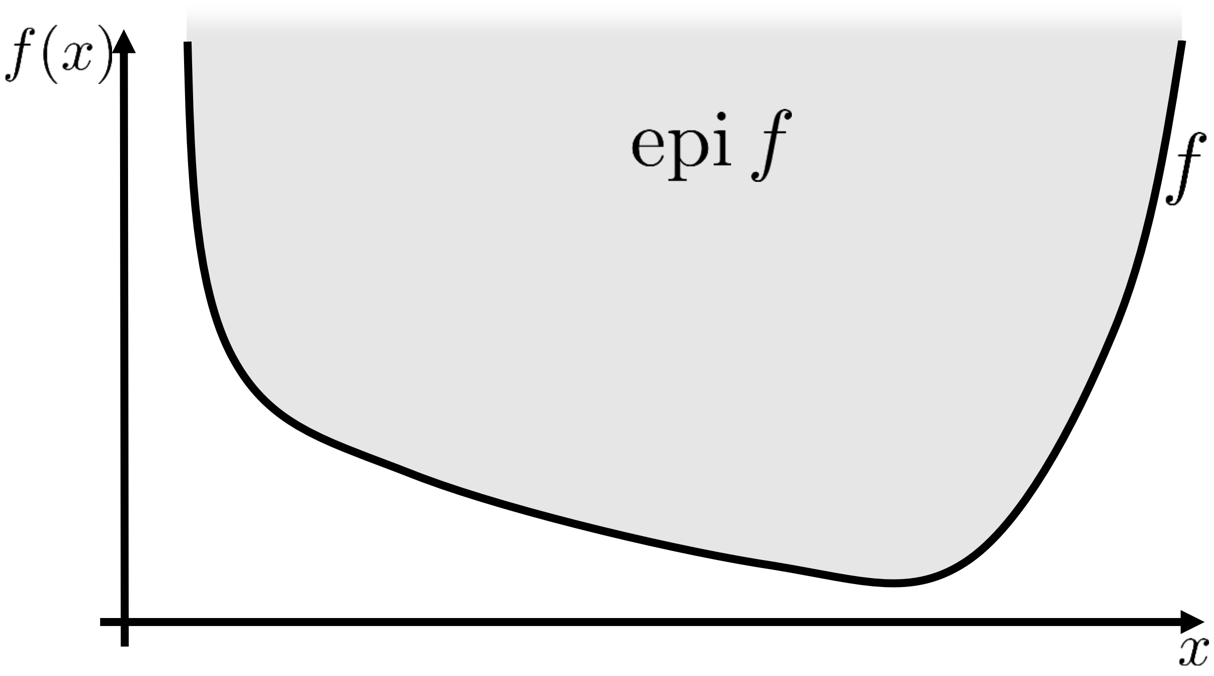
$$\operatorname{epi} f := \{(x,t) \mid x \in \mathcal{V}, t \in \mathbb{R}, f(x) \leq t\}$$

“Epi” is ancient greek and means “on top of”. The epigraph links the convexity property of functions with the one of sets. A function is said to be convex iff the epigraph of the function is a convex set. This also implies that the function domain needs to be a convex set.

Convexity



non-convex set



epigraph of a function

Convexity

Definition (Convex Function)

A function $f : \mathcal{V} \to \mathbb{R}$ is called *convex* if the function domain $\mathcal{V}$ is a convex set and if

$$f(\lambda x + (1-\lambda)y) \;\leq\; \lambda f(x) + (1-\lambda)f(y) \qquad \forall x, y \in \mathcal{V} \text{ and } 0 \leq \lambda \leq 1$$

The function is called *strictly convex* if the above inequality holds strictly for all $x, y \in \mathcal{V}$.
Further, a function f is called *(strictly) concave* if the function $-f$ is (strictly) convex.

The definition above is called the zero-order condition for convex functions. For (twice) differentiable functions corresponding first- and second-order conditions might sometimes be easier to check:

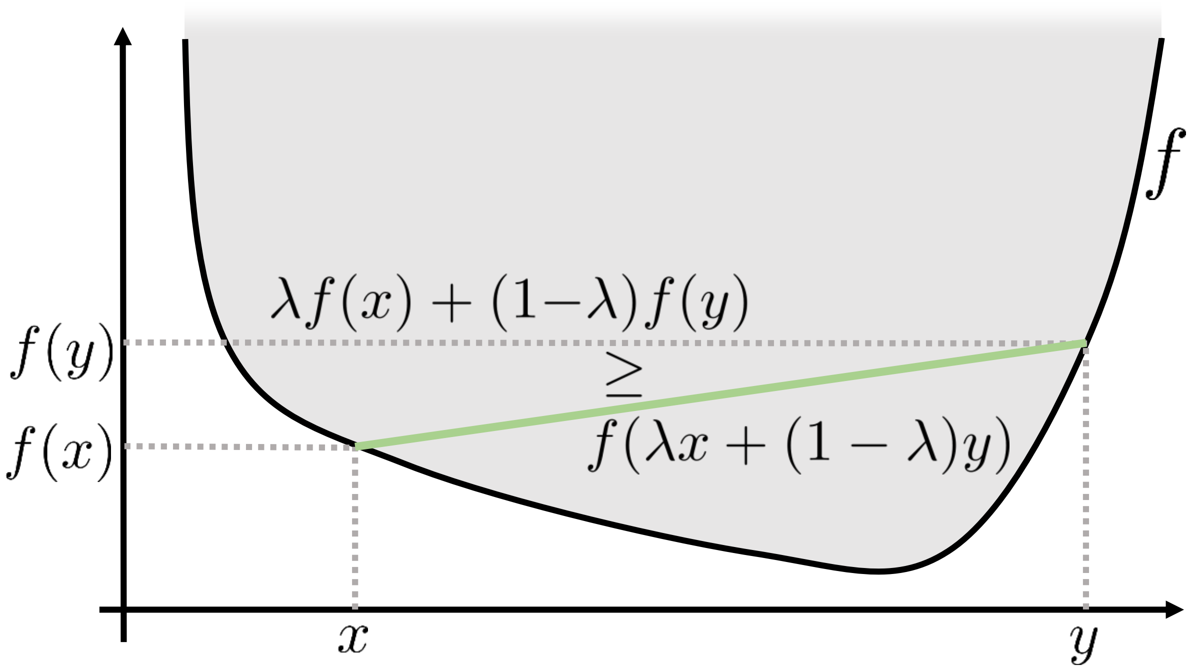
First order condition: $$f(y) \;\geq\; f(x) + \nabla f(x)^T (y - x) \qquad \forall x, y \in \mathcal{V}$$

"Function f is globally above the tangent at x."

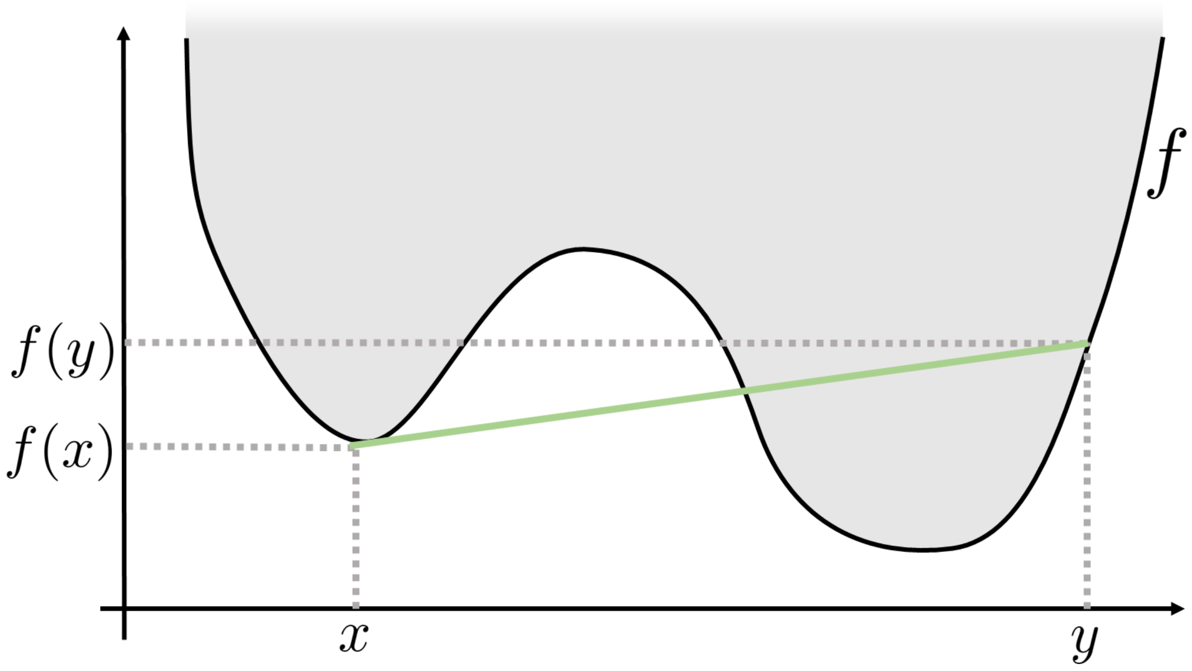
Second order condition: $$\nabla^2 f(x) \succeq 0 \qquad \forall x \in \mathcal{V}$$

"Function f is either flat or curved upwards in every direction."

Convexity



convex function

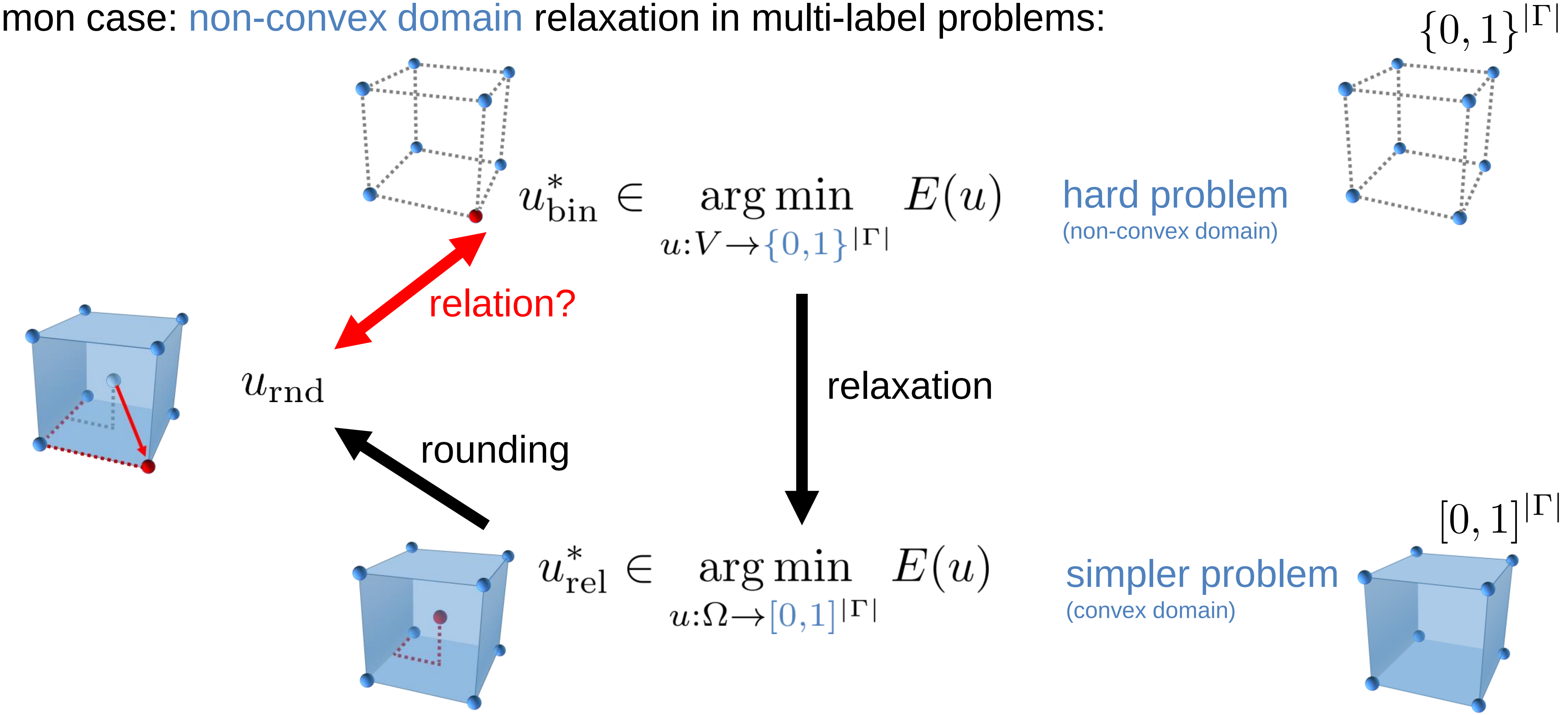


non-convex function

The most important and very useful property of convex functions is certainly the fact that any local minimum is also a global minimum.

Convex Relaxation

General idea: Make a non-convex problem convex, e.g. by relaxing constraints or by approximating the non-convex problem with a convex one.
Common case: non-convex domain relaxation in multi-label problems:



Rounding

Simplest rounding scheme: Locally select the most label, that is:

$$\forall x\colon \underbrace{u_{\text{rnd}}(x) = \hat{\mathbf{e}}_j}_{\text{only 1 label selected}} \text{ with } j = \underbrace{\min}_{\text{select min-index if several}} \underbrace{\left\{ \arg\max_{\ell} \left(u^*_{\text{rel}}(x) \right)_\ell \right\}}_{\text{get index with most likely label}} \text{ and } \hat{\mathbf{e}}_j \text{ being the } j\text{-th unit vector in } \{0,1\}^{|\Gamma|}.$$

Global optimality of the rounded solution with respect to the non-relaxed problem has been shown for the special case of binary labelings ($|\Gamma| = 1$) – see next Theorem.

Tightness of a relaxation:
In the case that solution of the relaxed problem has a deterministic and simple relationship (rounding) to the global optimal solution of the original problem, the relaxation is said to be tight.
The notion of tightness tries to assess the quality of a relaxation by means of energy values of the result or by means of a distance to the optimal relaxation (i.e. the convex envelope).

Rounding

Theorem (Thresholding Theorem – Equivalance of Minimizers) [Chan et al., SIAM JAM 2006]

Let functional $E : \mathcal{BV}(\Omega, [0,1]) \to \mathbb{R}$ be of the form $E(u) = \mathrm{TV}(u,\Omega) + \lambda \int_\Omega f u \, d\boldsymbol{x}$ with $f : \Omega \to \mathbb{R}$, $\lambda \in \mathbb{R}_{\geq 0}$ and let

$$u^*_{\mathrm{rel}} \in \underset{u \in \mathcal{BV}(\Omega,[0,1])}{\arg\min} \left\{ \mathrm{TV}(u,\Omega) + \lambda \int_\Omega f u \, d\boldsymbol{x} \right\}$$

be a global minimizer of the relaxed problem. Then, for any threshold value $\theta \in (0,1)$ the thresholded solution

$$u_{\mathrm{thr}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } u^*_{\mathrm{rel}}(\boldsymbol{x}) \geq \theta \\ 0 & \text{if } u^*_{\mathrm{rel}}(\boldsymbol{x}) < \theta \end{cases}$$

is a global minimizer of the corresponding binary minimization problem, that is

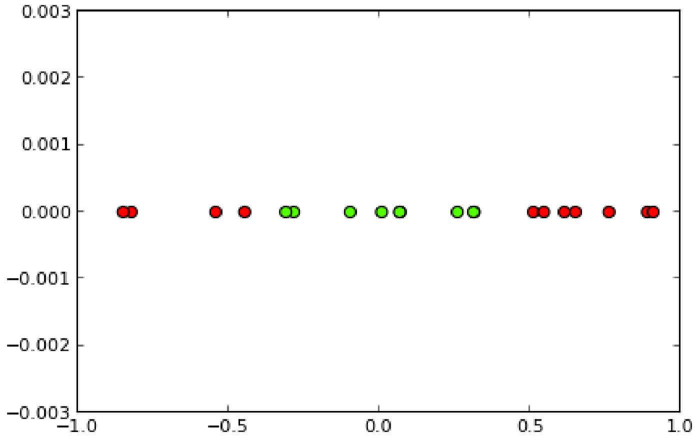
$$u_{\mathrm{thr}} = \mathbf{1}_{\{u^*_{\mathrm{rel}} \geq \theta\}} \in \underset{u \in \mathcal{BV}(\Omega,\{0,1\})}{\arg\min} \left\{ \mathrm{TV}(u,\Omega) + \lambda \int_\Omega f u \, d\boldsymbol{x} \right\}$$

Functional Lifting

General idea: Reformulate problem in a higher dimensional space to get a convex problem.

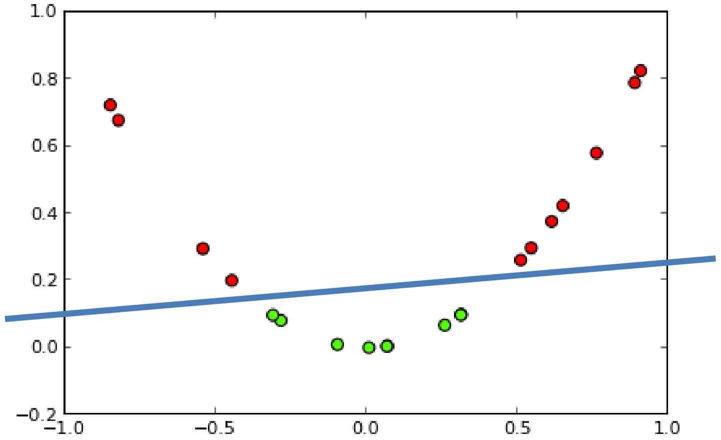
Compare to the lifting idea in linear classification:

$$f : \mathbb{R} \to \{0, 1\}, x \mapsto f(x)$$



Linear separation not possible

$$f : \mathbb{R} \times \mathbb{R} \to \{0, 1\}, (x, x^2) \mapsto f(x, x^2)$$



Linear separation possible after lifting input to higher dimension

Lifting: Multi-label Segmentation

[Chambolle et al., JIS 2012]

Given a set of labels, e.g. $\Gamma = \{1,2,3,4,5\}$ the labeling problem can be lifted from one integer dimension to multiple binary dimensions.

$$v^* \in \arg\min_{v:V\to\Gamma} E_v(v) \quad \xrightarrow[\text{Lifting}]{\text{Functional}} \quad u^* \in \arg\min_{u:V\to\{0,1\}^{|\Gamma|}} E_u(u)$$

e.g. $v^*(x) = 4$ e.g. $u^*(x) = (0,0,0,1,0)$

$$E(u) = \sum_{\ell=1}^{|\Gamma|} \int_\Omega \left[\lambda f_\ell(x) u_\ell(x) + \frac{1}{2}|\nabla u_\ell|_2\right] dx \qquad \text{s.t.} \sum_{\ell=1}^{|\Gamma|} u_\ell(x) = 1, \quad u(x) \geq 0 \qquad \forall x \in \Omega$$

Input Image

Segmentation with 10 regions

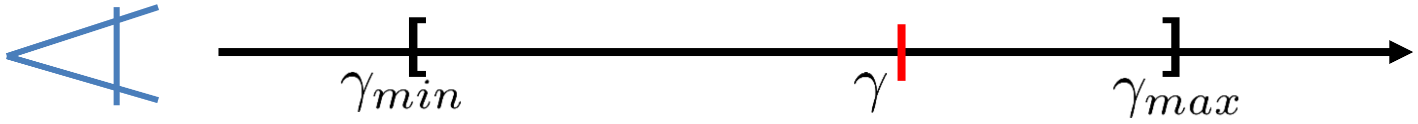
Lifting: Stereo Reconstruction

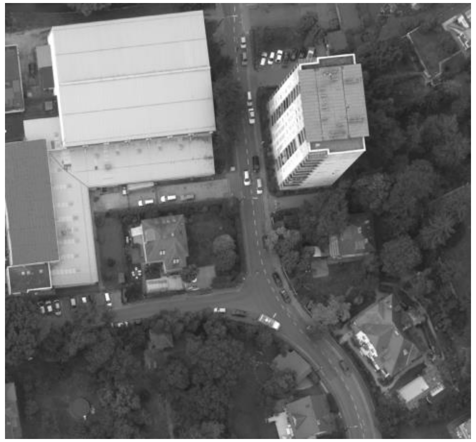
[Pock et al., ECCV 2008]

Given the label space of depth labels as $\Gamma = [\gamma_{min}, \gamma_{max}]$.

$$\min_{u:\Omega\to\Gamma} \int_\Omega \underbrace{|\nabla u(\boldsymbol{x})|\, d\boldsymbol{x}}_{\text{convex regularizer}} + \int_\Omega \underbrace{\rho(u(\boldsymbol{x}), \boldsymbol{x}) d\boldsymbol{x}}_{\text{non-convex matching term}}$$

viewing ray:





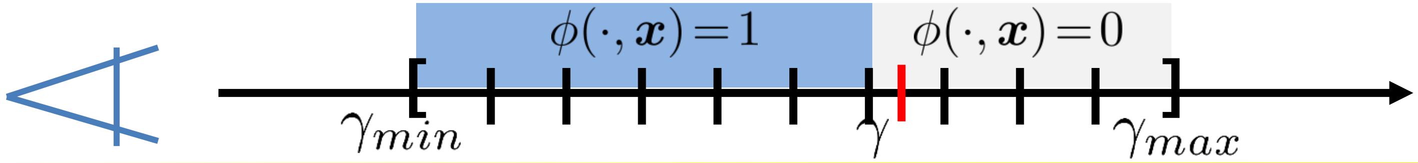
Input Image

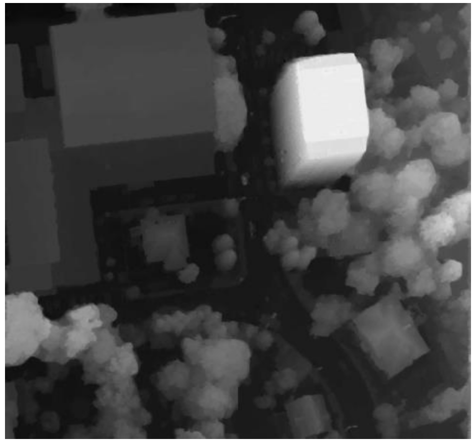
Lift problem to multi-dim. binary label space, define $\phi : [\Omega \times \Gamma] \to \{0, 1\}$ by $\phi(\boldsymbol{x}, \gamma) = \mathbf{1}_{\{u>\gamma\}}(x)$. Binary function now labels discretized depth as either free space or behind surface. Boundary conditions:

$$D = \{\phi : \Omega \times \Gamma \to \{0, 1\} \mid \phi(\gamma_{min}, \boldsymbol{x}) = 1, \phi(\gamma_{max}, \boldsymbol{x}) = 0\}$$

$$\min_{\phi\in D} \int_{\Omega\times\Gamma} \Big(\underbrace{|\nabla \phi(\gamma, \boldsymbol{x})|\, d\boldsymbol{x}}_{\text{convex regularizer}} + \underbrace{\rho(\gamma, \boldsymbol{x})\, |\partial_\gamma \phi(\gamma, \boldsymbol{x})|}_{\text{convex matching term}} \Big) d\gamma d\boldsymbol{x}$$

viewing ray:





Depth Labeling

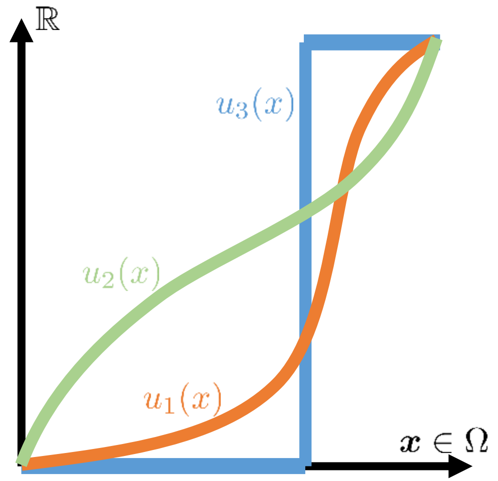
Total Variation

Definition (Total Variation for differentiable functions)

Let $u \in \mathcal{C}^1(\Omega, \mathbb{R}^n)$ be a differentiable function, then the functional

$$\mathrm{TV}(u,\Omega) = \int_{\Omega} |\nabla u|_2 \; d\boldsymbol{x}$$

is called the total variation of function u on the domain Ω.



3 functions with equal TV

Definition (Function space of bounded variation)

The set of functions with bounded variation, that is, with a variation smaller than infinity, is defined as

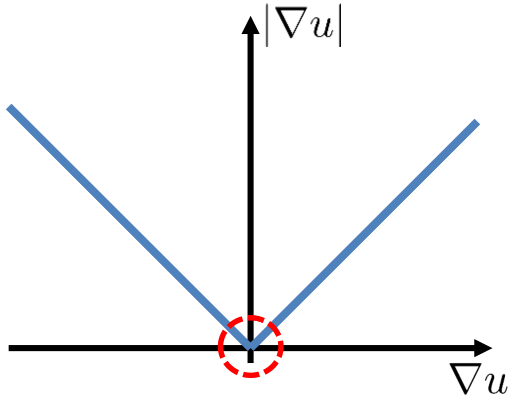
$$\mathcal{BV}(\Omega,\mathbb{R}) = \left\{ u : \Omega \to \mathbb{R} \;\middle|\; \mathrm{TV}(u,\Omega) < \infty \right\}$$

The total variation defines a semi-norm on the space of bounded variations. For simplicity it is often called TV-norm.

Total Variation

Some properties of the total variation, it:

- is a measure of smoothness.
- is a convex functional (proof is part of the homework)
- is a so-called “robust” norm, because it does not over-penalize outliers
- is the best trade-off between being robust to outliers and efficient to optimize (because of convexity)
- is a suitable regularizer to preserve discontinuities in the signal (because one large gradient has the same cost as many small ones that finally reach the same function value)
- is not differentiable at locations with zero gradient



Weighted Total Variation

Definition (Weighted Total Variation for differentiable functions) [Bresson et al., JMIV 2009]

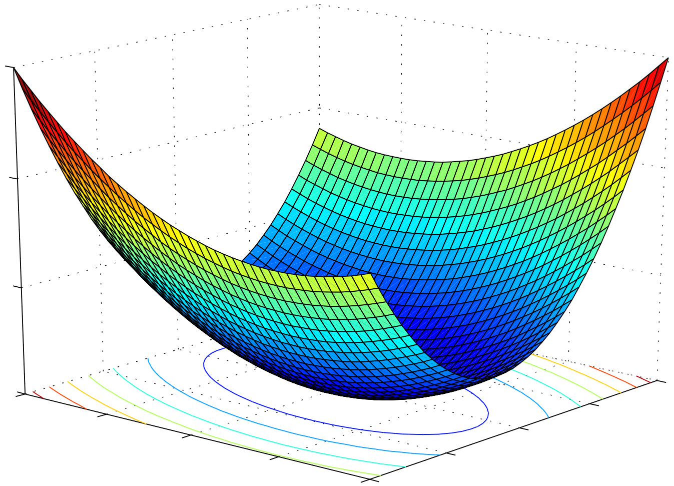
Let $u \in \mathcal{C}^1(\Omega, \mathbb{R}^n)$ be a differentiable function and $g : \Omega \to \mathbb{R}_{\geq 0}$ be a weight function, then the functional

$$\mathrm{TV}_g(u, \Omega) = \int_\Omega g(\boldsymbol{x}) \, |\nabla u|_2 \; d\boldsymbol{x}$$

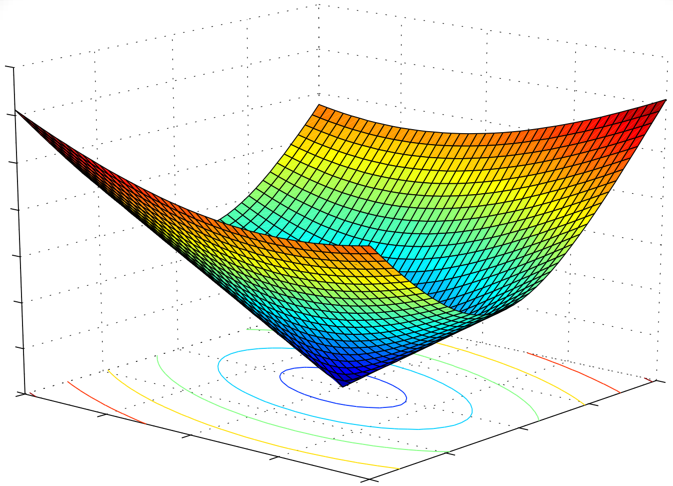
is called the weighted total variation of function u on the domain Ω.

Bresson et al. showed that the corresponding generalized version of the thresholding theorem still holds when an additional weight is introduced to the total variation term.

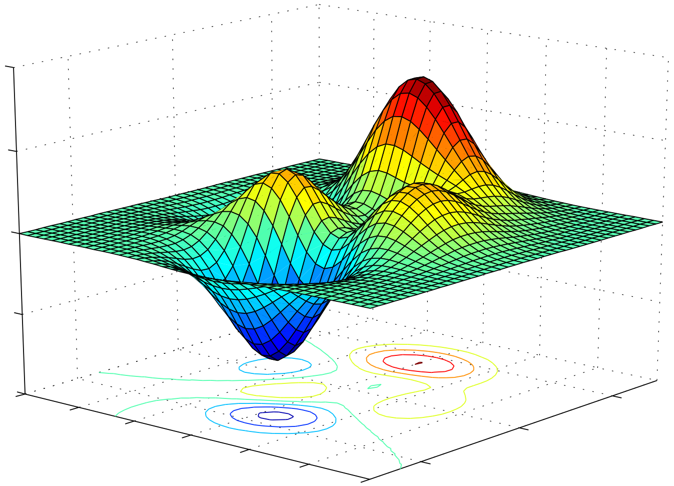
Non-Smooth Convex Optimization



smooth convex



non-smooth convex



non-convex

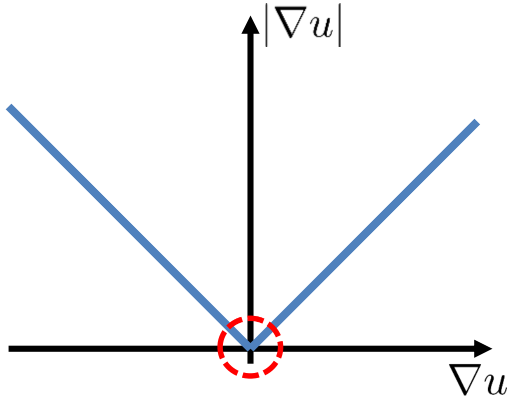
TV-Optimization Algorithms

Many algorithms available to minimize Total Variation energies (only highlighted ones are discussed):

- (Projected) Gradient Descent
- Lagged Diffusivity Fixed Point Iterations (LDFPI)
- Fast Iterative Shrinkage Algorithm (FISTA)
- Alternating Direction Method of Multipliers (ADMM)
- First-order Primal-Dual Algorithm
- …

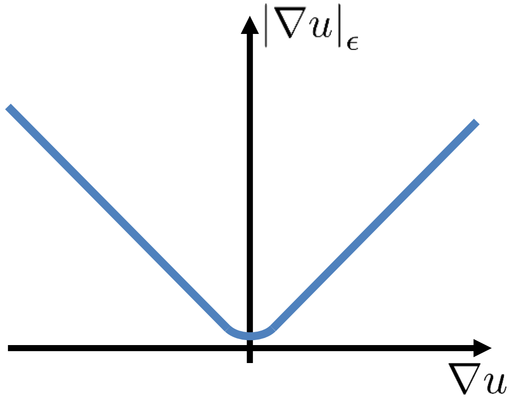
TV-Norm Pertubation

Problem of optimizing TV-functionals: non-differentiability at zero



One possibility to deal with this problem was proposed by [Rudin, Osher, Fatemi, 1992]. They proposed to use a perturbation of the TV-norm:

$$\mathrm{TV}(u,\Omega)_\epsilon = \int_\Omega |\nabla u|_\epsilon \, d\boldsymbol{x} \qquad \text{with} \quad |\nabla u|_\epsilon = \sqrt{|\nabla u|_2^2 + \epsilon}$$



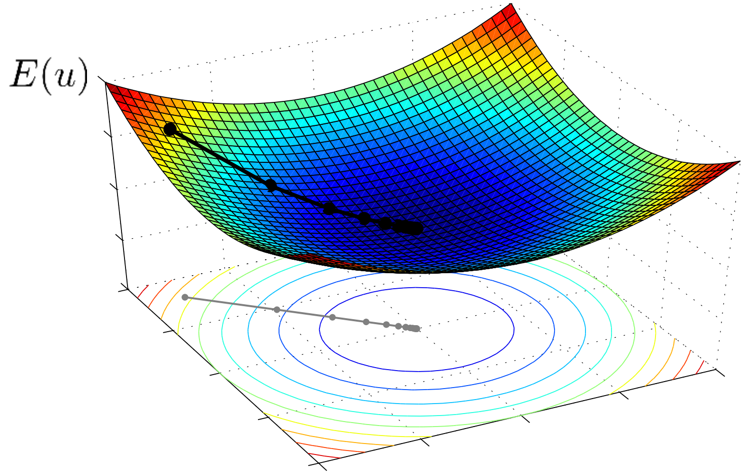
where $\epsilon > 0$ is a small positive number.

Gradient Descent

Gradient Descent

$$\frac{du}{dt} = -\frac{dE}{du}$$

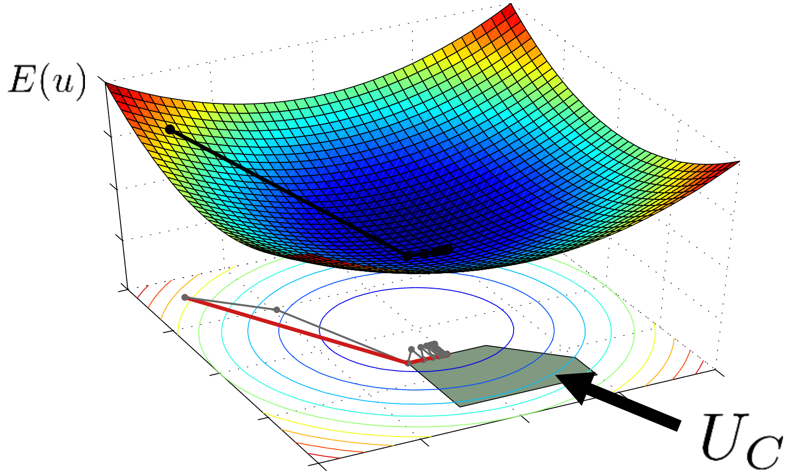
$$u^{k+1} = u^k - \tau \frac{dE(u^k)}{du}$$



gradient descent

Projected Gradient Descent

$$u^{k+1} = \Pi_{U_C}\left[u^k - \tau \frac{dE(u^k)}{du}\right]$$



projected gradient descent

Duality

Definition (Convex Conjugate – Legendre-Fenchel Transform)

Let $f : \mathcal{V} \to \mathbb{R}$ be a function. Then, the function $f^* : \mathcal{V}^* \to \mathbb{R}$,

$$f^*(\boldsymbol{p}) = \sup_{\boldsymbol{x} \in \mathcal{V}} \{\langle \boldsymbol{x}, \boldsymbol{p} \rangle - f(\boldsymbol{x})\}$$

is called the convex conjugate or Legendre-Fenchel transform of the function f and $\mathcal{V}^*$ is called the dual space of $\mathcal{V}$.

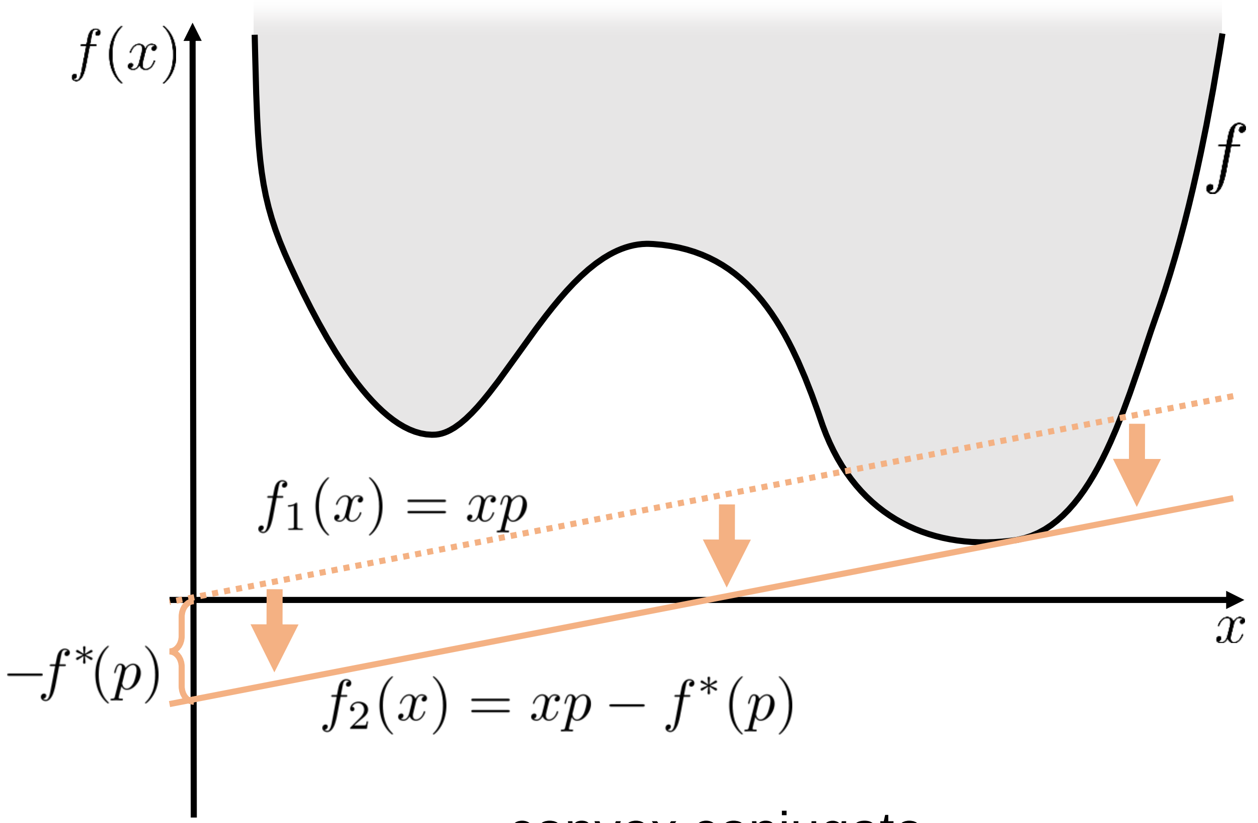
The idea behind the Legendre-Fenchel transform is to represent the function f in the space of supporting lines (or hyperplanes) of the graph being represented as tuples of the slope (or plane normal) and the corresponding maximal intercept. Hence, the supremum operation is used to transform from the space of $\left(\boldsymbol{x}, f(\boldsymbol{x})\right)$ to the space of gradient and intercept $\left(\boldsymbol{p}, f^*(\boldsymbol{p})\right)$ of the tangent.

The convex biconjugate $f^{**} = (f^*)^* \leq f$ is the maximal convex function below f and represents the convex hull of the epigraph of f. It is also called the convex envelope of function f.

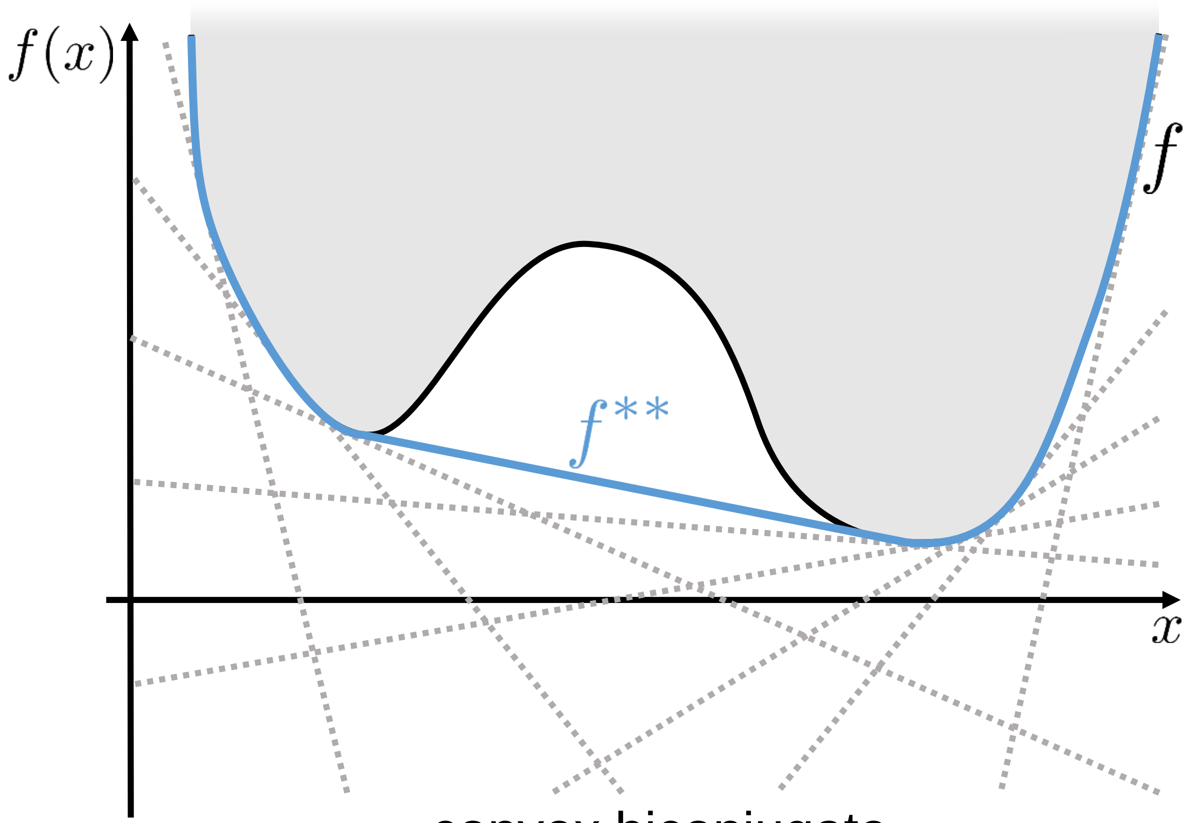
Duality

$$f^*(\boldsymbol{p}) = \sup_{\boldsymbol{x}\in\mathcal{V}} \{\langle \boldsymbol{x}, \boldsymbol{p}\rangle - f(\boldsymbol{x})\}$$

$$f^{**}(\boldsymbol{p}) = \sup_{\boldsymbol{p}\in\mathcal{V}^*} \{\langle \boldsymbol{x}, \boldsymbol{p}\rangle - f^*(\boldsymbol{p})\}$$



convex conjugate



convex biconjugate

Note: $f^*(p)$ is the intercept of the tangent with slope p.

Duality

Definition (Adjoint Operator)

Let $\mathcal{H}_x, \mathcal{H}_y$ be Hilbert spaces, with respective inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_x}, \langle \cdot, \cdot \rangle_{\mathcal{H}_y}$ and further let $A : \mathcal{H}_x \to \mathcal{H}_y$ be a continuous linear operator. One can show that there exists a unique continuous linear operator $A^* : \mathcal{H}_y \to \mathcal{H}_x$ with the following property:

$$\langle Ax, y \rangle_{\mathcal{H}_y} = \langle x, A^* y \rangle_{\mathcal{H}_x} \qquad \forall x \in \mathcal{H}_x, y \in \mathcal{H}_y.$$

Operator A^* is called the adjoint of A.

Duality

Definition (Weak Derivative)

Let $\Omega \subset \mathbb{R}^n$ and $u \in \mathcal{L}^2(\Omega)$, then function $v \in \mathcal{L}^2(\Omega)$ is a weak derivative of u if,

$$\int\limits_{\Omega} u \cdot \operatorname{div}(\boldsymbol{p})\, d\boldsymbol{x} = -\int\limits_{\Omega} v \cdot \boldsymbol{p}\, d\boldsymbol{x}$$

for all functions $\boldsymbol{p}$ being infinitely differentiable and with compact support in Ω, i.e. $\boldsymbol{p} \in \mathcal{C}_c^{\infty}(\Omega, \mathbb{R}^n)$

This equation can be derived via the "integration by parts" method. The corresponding third boundary term vanishes due to the required compact support property, which essentially means that all test functions $\boldsymbol{p}$ are zero at the domain boundary $\partial\Omega$. The key idea is to "shift" the differential operator to another variable, which is defined to be differentiable everywhere. Everywhere where u has a classical (strong) derivative it holds that $v(\boldsymbol{x}) = \nabla u(\boldsymbol{x})$.
Note that both sides of the equation correspond to an inner product in the function space, e.g. $\int u \cdot \operatorname{div}(\boldsymbol{p})\, d\boldsymbol{x} = \langle u, \operatorname{div}(\boldsymbol{p})\rangle$. With the definition of the adjoint operator, we can derive that the differential operators ∇u and $-\operatorname{div}(\boldsymbol{p})$ are adjoint.

Duality

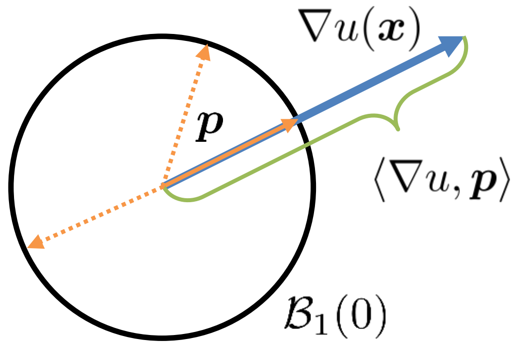
Property of the 2-norm:

$$|\nabla u|_2 = |\nabla u|_2 \cdot \frac{|\nabla u|_2}{|\nabla u|_2} = \frac{\langle \nabla u, \nabla u \rangle}{|\nabla u|_2} = \langle \nabla u, \underbrace{\frac{\nabla u}{|\nabla u|_2}}_{\boldsymbol{p}} \rangle \quad \text{for } \nabla u \neq 0$$

Leads to the dual representation of the 2-norm:

$$|\nabla u(\boldsymbol{x})|_2 = \sup_{\|\boldsymbol{p}(\boldsymbol{x})\|_2 \leq 1} \langle \nabla u(\boldsymbol{x}), \boldsymbol{p}(\boldsymbol{x}) \rangle$$

Unit ball at zero:



$$\langle \nabla u, p \rangle = \|\nabla u\| \underbrace{\|p\|}_{1} \underbrace{\cos(\theta)}_{1}$$

Total Variation

Definition (Total Variation) [Chan et al., SIAM JSC 1999]

A functional of the form

$$\mathrm{TV}(u,\Omega) := \sup\left\{ -\int_\Omega u \cdot \mathrm{div}(\boldsymbol{p})\, d\boldsymbol{x} \;\middle|\; \boldsymbol{p} \in \mathcal{C}_c^1(\Omega,\mathbb{R}^n), \|\boldsymbol{p}\|_{\mathcal{L}^\infty(\Omega)} \le 1 \right\}$$

is called the total variation of function u on the domain Ω.

This definition of TV is a generalization of the classical term $\int |\nabla u|_2 \; d\boldsymbol{x}$ to weakly differentiable functions.

Often in papers the classical definition is given, but the authors mean the more general definition given above - sometimes with the side note that the derivative is meant in a “distributional” way which is sometimes also written as $\int |Du|_2 \; d\boldsymbol{x}$, which is then called the “distributional derivative”.

Total Variation

The generalization can be verified by applying a differentiable function $u \in \mathcal{C}^1(\Omega, \mathbb{R})$ and using the vector field $\boldsymbol{p} \in \mathcal{L}^1(\Omega, \mathbb{R}^n)$ defined by

$$\boldsymbol{p}(\boldsymbol{x}) = \begin{cases} \frac{\nabla u(\boldsymbol{x})}{|\nabla u(\boldsymbol{x})|_2} & \text{if } |\nabla u(\boldsymbol{x})|_2 \neq 0 \\ 0 & \text{otherwise} \end{cases},$$

then previous definition of the TV reduces to the classical definition $\int |\nabla u|_2 \, d\boldsymbol{x}$.

Definition (Weighted Total Variation) [Bresson et al., JMIV 2009]

A functional of the form

$$\mathrm{TV}_g(u, \Omega) := \sup \left\{ -\int_\Omega u \cdot \mathrm{div}(\boldsymbol{p}) \, d\boldsymbol{x} \;\middle|\; \boldsymbol{p} \in \mathcal{C}_c^1(\Omega, \mathbb{R}^n), \|\boldsymbol{p}\|_{\mathcal{L}^\infty(\Omega)} \leq g(x) \right\}$$

is called the weighted total variation of function u on the domain Ω.

Primal-Dual Algorithm

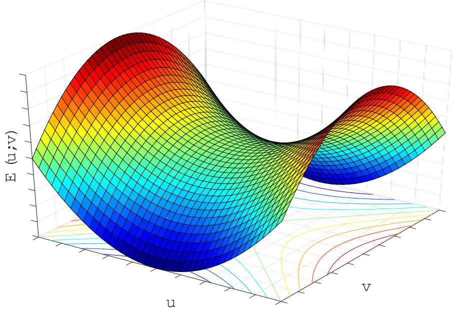
[Pock et al., ICCV 2009]
[Chambolle, Pock., JMIV 2011]

The first-order primal-dual Algorithm minimizes energies of the following canonical form:

$$\min_{x \in X} F(Kx) + G(x)$$

with $F, G : X \to \mathbb{R}$ being proper, convex, lower-semicontinuous functions and $K : X \to Y$ being a linear operator. Function F can be non-differentiable. A common case is that F represents a regularizer in form of a norm and $K = \nabla$, $K^* = -\operatorname{div}$ are the gradient operators.
Using the definitions of Legendre-Fenchel transform and the adjoint operator, the energy can be transformed as follows:

$$
\begin{aligned}
&\min_{x \in X} F(Kx) + G(x) && \text{(primal)} \\
&= \min_{x \in X} \max_{y \in Y} \langle Kx, y \rangle - F^*(y) + G(x) && \text{(primal-dual)} \\
&= \max_{y \in Y} \min_{x \in X} \langle x, K^* y \rangle - F^*(y) + G(x) && \text{(dual-primal)} \\
&= \max_{y \in Y} -(F^*(y) + G^*(-K^* y)) && \text{(dual)}
\end{aligned}
$$



Saddle Point Problems

Primal-Dual Algorithm

[Pock et al., ICCV 2009]
[Chambolle, Pock., JMIV 2011]

Primal Dual algorithm

- Initialize: $x^0 \in X, y^0 = 0, \bar{x}^0 = x^0$
- Iterate:

$$y^{k+1} = \mathrm{prox}_{\sigma F^*}\left(y^k + \sigma K \bar{u}^k\right)$$

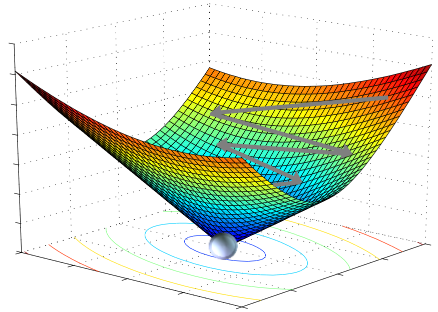
(projected) gradient ascent in the dual variable

$$x^{k+1} = \mathrm{prox}_{\tau G}\left(x^k - \tau K^* y^{k+1}\right)$$

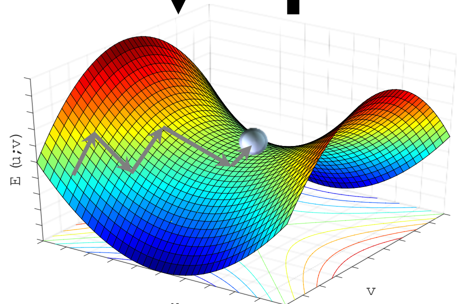
(projected) gradient descent in the primal variable

$$\bar{x}^{k+1} = x^{k+1} + \theta(x^{k+1} - x^k)$$

linear extrapolation step of the primal variable



non-smooth convex



(smooth) saddle

For suitable step sizes $\sigma > 0, \tau > 0, \tau\sigma L^2 < 1$ and $\theta = 1$ the algorithm has been shown to converge, where $L = \|K\|$ is the Lipschitz constant.

The algorithm minimizes the gap between primal and dual energies:

$$\mathrm{Gap}(x, y) = F(Kx) + G(x) + F^*(y) + G^*(-K^*y)$$

ETH zürich

Primal-Dual Algorithm

[Chambolle, Pock., JMIV 2011]

The proximity or prox operator $\mathrm{prox}_{\tau G}(u)$, also the resolvent operator $(I + \tau\partial G)^{-1}(u)$ is a generalized projection and is defined as

$$\mathrm{prox}_{\tau G}(u) := \arg\min_{v} \left\{ \frac{1}{2} \|u - v\|^2 + \tau G(v) \right\}$$

For the special case that G is an indicator function of a convex set C, that is, $G(x) = \chi_C(x)$ with $\chi_C(x) := \{0 \text{ if } x \in C, \infty \text{ else}\}$, then the prox operator simplifies to the Euclidean projection, denoted as, Π_C onto the set C. Thus, $\mathrm{prox}_{\chi_C}(x) = \Pi_C(x)$.

Preconditioning

The idea of preconditioning is to make the energy landscape more isotropic, to reduce the effect of zig-zag patterns during gradient-based optimization schemes. The amount of isotropy is described by the condition number.

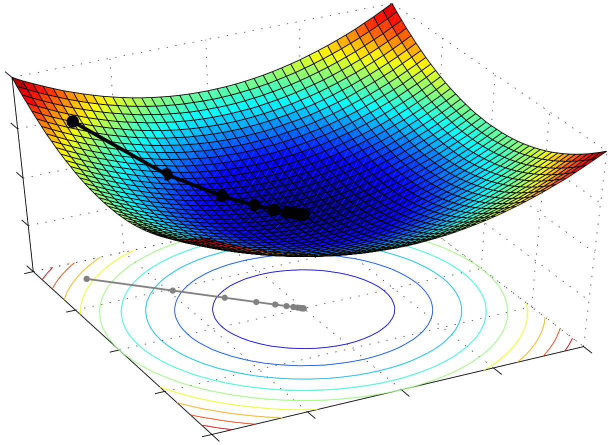
Example: Linear system solver

$$Ax = b \quad \Rightarrow \quad AP^{-1}Px = b$$

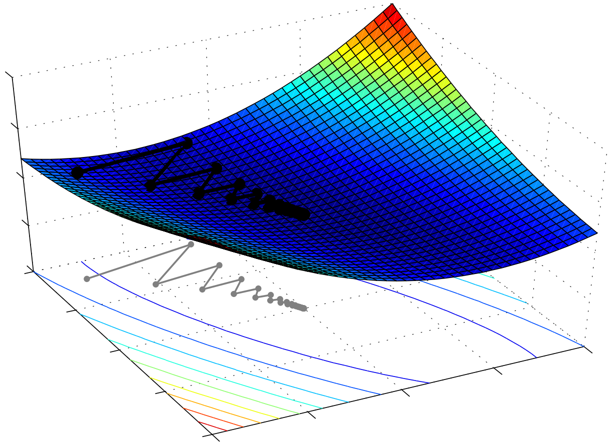
$$\Rightarrow \quad AP^{-1}y = b \text{ and } Px = y$$

condition number

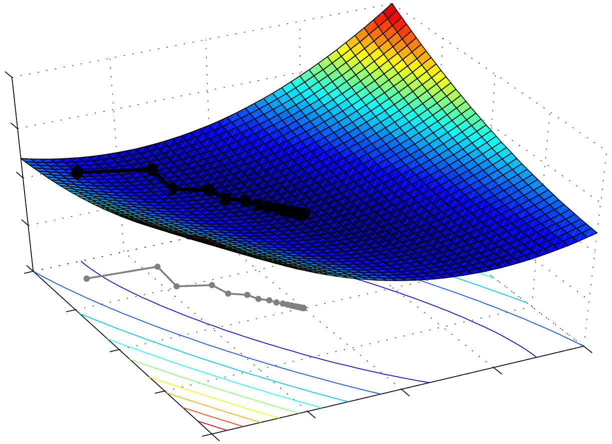
$$\kappa(A) = \left| \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \right|$$



Isotropic function



anisotropic function
without step-size adaption



anisotropic function
with step-size adaption

Preconditioned Primal-Dual Algorithm

[Pock, Chambolle, ICCV 2011]

Preconditioned Primal Dual algorithm

- Initialize: $x^0 \in X, y^0 = 0, \bar{x}^0 = x^0$
- Iterate:

$$y^{k+1} = \text{prox}_{\Sigma F^*}\left(y^k + \Sigma K \bar{u}^k\right)$$

(projected) gradient ascent in the dual variable

$$x^{k+1} = \text{prox}_{TG}\left(x^k - TK^* y^{k+1}\right)$$

(projected) gradient descent in the primal variable

$$\bar{x}^{k+1} = x^{k+1} + \theta(x^{k+1} - x^k)$$

linear extrapolation step of the primal variable

The global step sizes σ, τ are replaced by individual steps for each dimension, defined by the diagonal matrices $T = \text{diag}(\tau_1, \dots, \tau_n), \Sigma = \text{diag}(\sigma_1, \dots, \sigma_m)$ with

$$\tau_j = \frac{1}{\sum_{i=1}^{m} |K_{i,j}|^{2-\alpha}} \quad , \qquad \sigma_i = \frac{1}{\sum_{j=1}^{n} |K_{i,j}|^{\alpha}}$$

for any $\alpha \in [0, 2]$, typically $\alpha = 1$, where $m \times n$ are the dimensions of K.

Convergence Criteria

- Fixed number of iterations

- Bound on energy change

$$\left|\frac{E(u^{k-1}) - E(u^k)}{E(u^k)}\right| < \theta_E$$

- Bound on Primal-Dual Gap change

$$\left|\frac{\mathrm{Gap}(u^{k-1}, \boldsymbol{p}^{k-1}) - \mathrm{Gap}(u^k, \boldsymbol{p}^k)}{\mathrm{Gap}(u^k, \boldsymbol{p}^k)}\right| < \theta_{\mathrm{Gap}}$$

- Bound on solution change

$$\frac{|u^k - u^{k+1}|}{|u^k|} < \theta_U$$

In practice, a combination of above convergence criteria can be an effective trade-off, e.g. check the convergence every 100 iterations and define a maximum fixed number of iterations.

Discretization

The image domain is discretized on a regular Cartesian grid:

$$\Omega = \left\{ (i \cdot h_\Omega, j \cdot h_\Omega) \in \mathbb{R}^2 \,\middle|\, i, j \in \mathbb{Z},\ 1 \le i \le N,\ 1 \le j \le M \right\}$$

The gradient $\nabla u = \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right)^T$ is discretized as $(\nabla u)_{i,j} = \left(\delta_x^+ u_{i,j}, \delta_y^+ u_{i,j}, \right)^T$ and

discretize derivatives for the gradient operator are approximated by forward differences and Neumann boundary conditions:

$$\delta_x^+ u_{i,j} = \frac{u_{i+1,j} - u_{i,j}}{h_\Omega} \qquad \delta_y^+ u_{i,j} = \frac{u_{i,j+1} - u_{i,j}}{h_\Omega}$$

Due to adjointness the divergence operator $\operatorname{div} \boldsymbol{p} = \frac{\partial p^1}{\partial x} + \frac{\partial p^2}{\partial y}$ of a vector field $\boldsymbol{p} = (p^1, p^2)$ is discretized with corresponding backward differences:

$$(\operatorname{div} \boldsymbol{p})_{i,j} = \delta_x^- p^1_{i,j} + \delta_y^- p^2_{i,j}$$ defined as $$\delta_x^- p_{i,j} = \frac{p^1_{i,j} - p^1_{i-1,j}}{h_\Omega} \qquad \delta_y^- p_{i,j} = \frac{p^2_{i,j} - p^2_{i,j-1}}{h_\Omega}$$

References

- R. Tyrrell Rockafellar, “*Convex Analysis*”, Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970
- R. Tyrrell Rockafellar and Roger J-B Wets, “*Variational Analysis*”, volume 317 of Grundlehren der Mathematischen Wissenschaften. Springer, 1998
- Stephen Boyd and Lieven Vandenberghe, “*Convex Optimization*”, Cambridge University Press, 2004
- L. Ambrosio, N. Fusco, and D. Pallara, “*Functions of bounded variation and free discontinuity problems*”, Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, 2000
- Tony F. Chan, Gene H. Golub, and Pep Mulet, “*A nonlinear primal-dual method for total variation-based image restoration.*”, SIAM J. Sci. Comput., 20(6):1964–1977, 1999
- T. Chan, S. Esedoglu, and M. Nikolova, “*Algorithms for finding global minimizers of image segmentation and denoising models.*”, SIAM Journal on Applied Mathematics, 66(5):1632–1648, 2006
- Xavier Bresson, Selim Esedoglu, Pierre Vandergheynst, Jean-Philippe Thiran, and Stanley Osher, “*Fast global minimization of the active contour/snake model.*”, Journal of Mathematical Imaging and Vision, 28(2):151–167, 2007
- T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers, “*A Convex Formulation of Continuous Multi-label Problems.*”, In Proc. International Conference on Computer Vision (ICCV), Kyoto, Japan, 2009
- T. Pock, D. Cremers, H. Bischof, and A. Chambolle, “*An algorithm for minimizing the piecewise smooth Mumford-Shah functional.*”, In ECCV, 2008.
- Antonin Chambolle and Thomas Pock, “*A first-order primal-dual algorithm for convex problems with applications to imaging*”, Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011
- Thomas Pock and Antonin Chambolle, “*Diagonal preconditioning for first order primal- dual algorithms in convex optimization*”, ICCV, 2011
- M. Klodt, T. Schoenemann, K. Kolev, M. Schikora, D. Cremers, “*An Experimental Comparison of Discrete and Continuous Shape Optimization Methods*”, European Conference on Computer Vision (ECCV), 2008
- A. Chambolle, D. Cremers, and T. Pock, “*A convex approach to minimal partitions.*”, SIAM Journal on Imaging Sciences, 5(4):1113–1158, 2012

Slide credits: Several slides are direct copies / are inspired by, or use material from the lecture slides from Prof. Bastian Goldlücke.