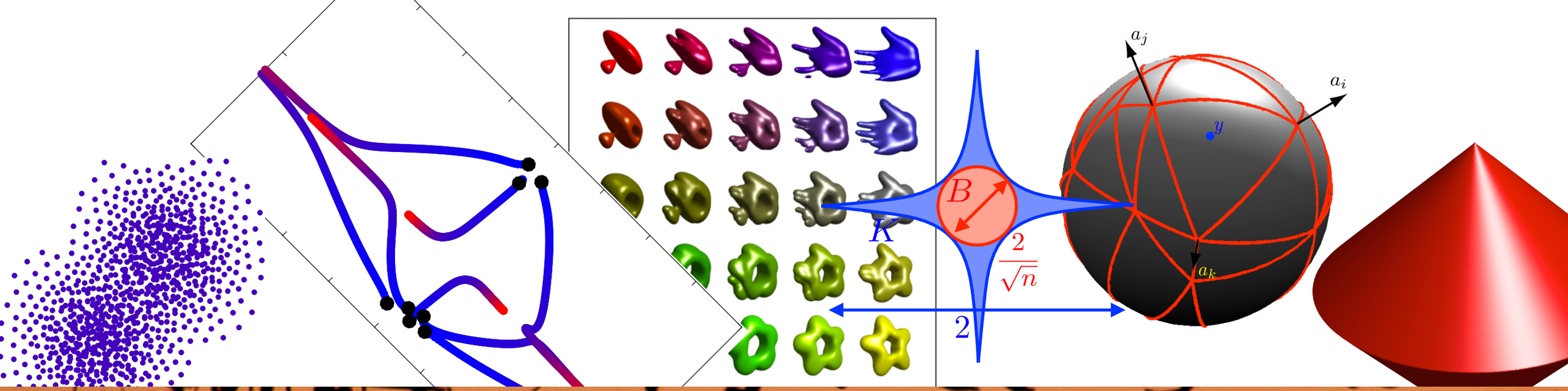
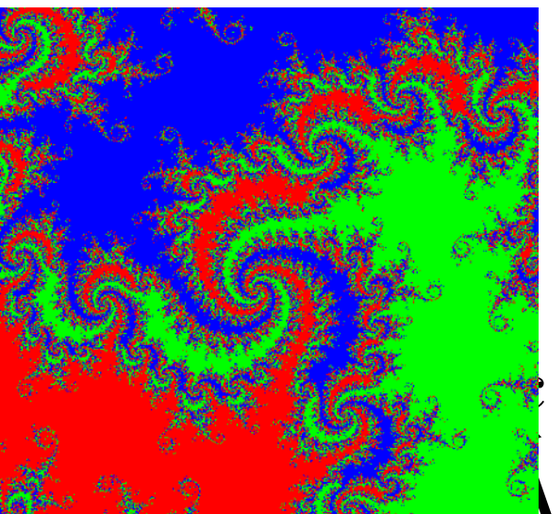
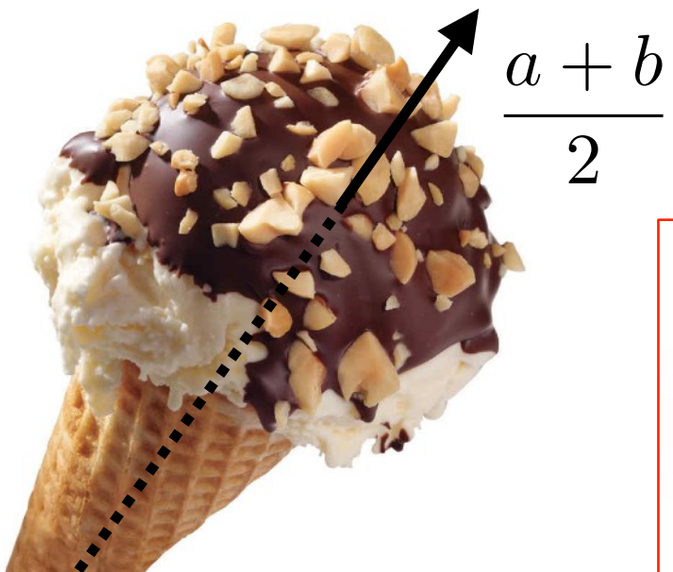
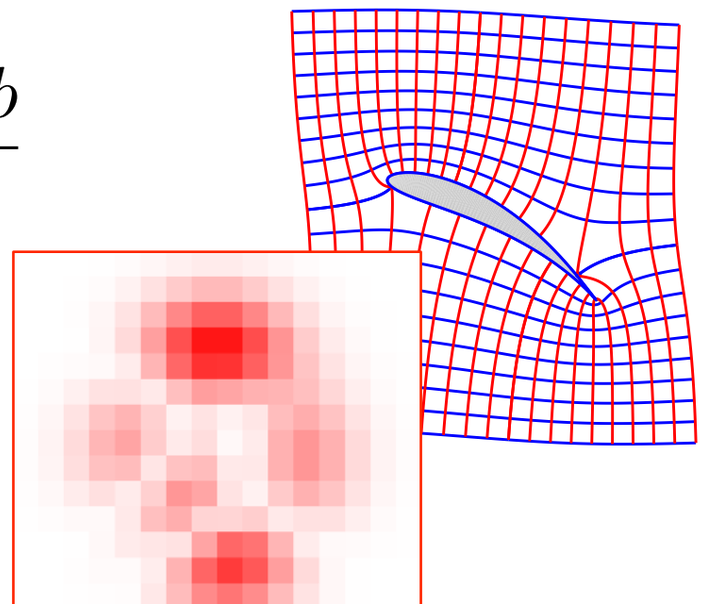


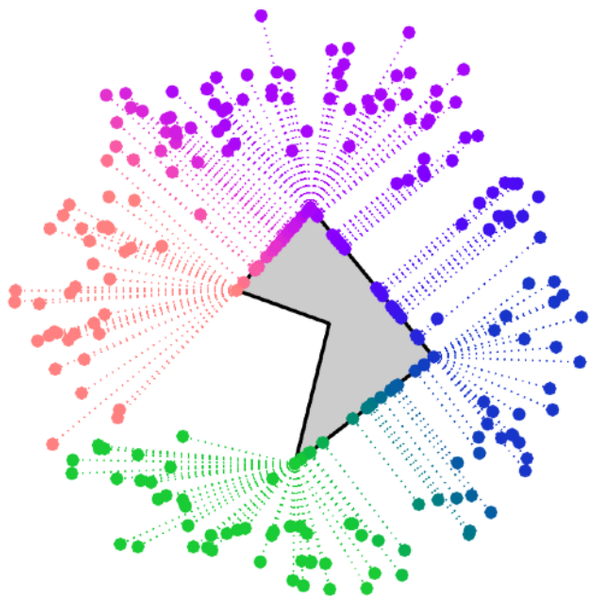
Tweets

slides and code









Optimization

$$x^\star \in \underset{x \in X}{\operatorname{argmin}}\ f(x)$$

$$\iff \quad \delta_{x^\star} \in \underset{\mu \text{ prob. distrib.}}{\operatorname{argmin}} \int_X f(x) \mathrm{d}\mu(x)$$

Any optimization program is a equivalent to a convex (linear) one. #RadonMeasuresRule

$$f^*(\omega) \stackrel{\text{def.}}{=} \sup_x \langle x, \omega \rangle - f(x)$$

$$(f \diamond g)(x) \stackrel{\text{def.}}{=} \inf_y f(y) + g(x-y)$$

$$\textit{Theorem: } (f \diamond g)^* = f^* + g*$$

$$\hat{f}(\omega) \stackrel{\text{def.}}{=} \int f(x) e^{-\mathrm{i}\langle \omega,\, x \rangle} \mathrm{d}x$$

$$(f \star g)(x) \stackrel{\text{def.}}{=} \int f(y) g(x-y) \mathrm{d}y$$

$$\textit{Theorem: } \widehat{f \star g} = \hat{f} \cdot \hat{g}$$

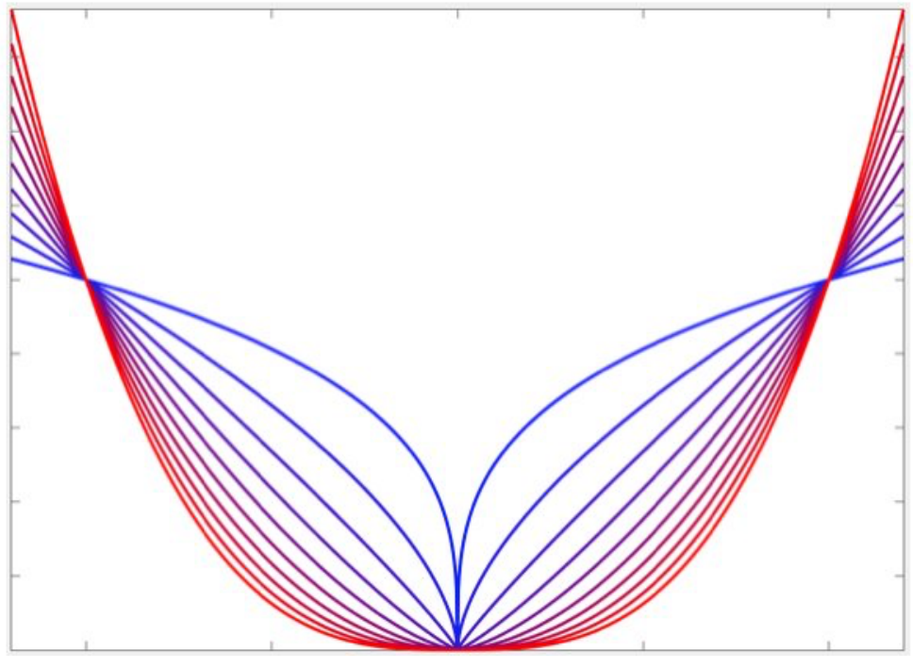
Fourier is to convolution what Legendre is to inf-convolution.
https://en.wikipedia.org/wiki/Fourier_transform#Convolution_theorem …
https://en.wikipedia.org/wiki/Legendre_transformation#Infimal_convolution …

Kurdyka-Łojasiewicz:

(at minimum $f(0) = 0$)

$$\exists \varphi, \|\nabla(\varphi \circ f)\| \geqslant 1$$

$$\iff \quad \exists (K, \tau), \|\nabla f(x)\| \geqslant K|f(x)|^{\tau}$$



Kurdyka-Łojasiewicz property: fundamental to show convergence of descent optimization methods. tau controls rate. https://arxiv.org/abs/0802.0826v1

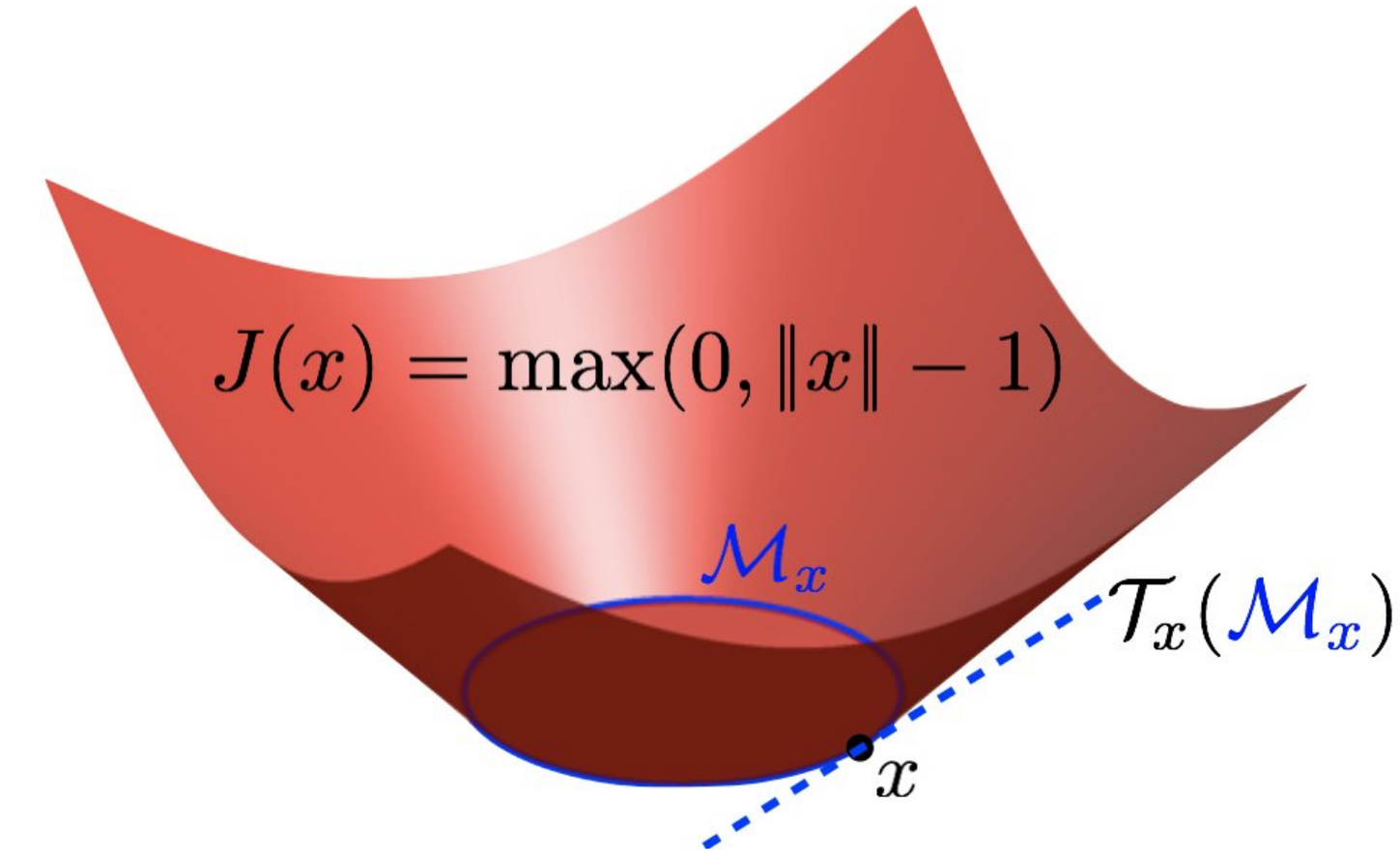
$J : \mathbb{R}^N \to \mathbb{R}$ is partly smooth at x for a manifold $\mathcal{M}_x$

(i) J is C^2 along $\mathcal{M}_x$ around x ;

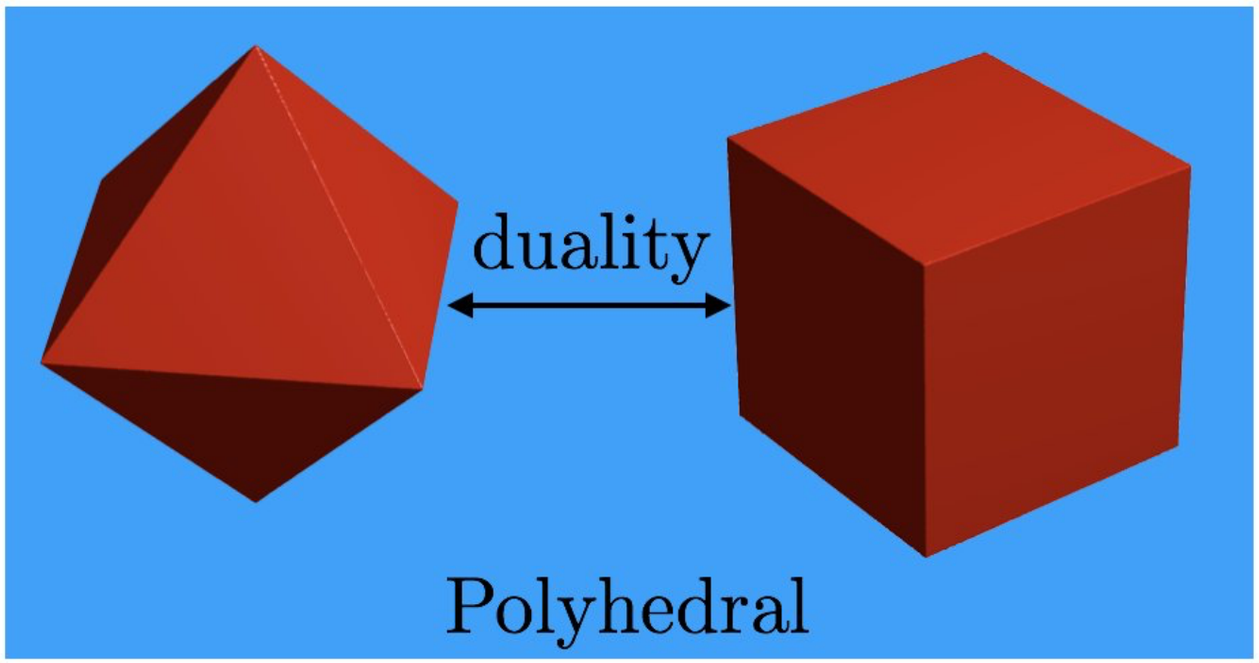
(ii) $\forall\, h \in \mathcal{T}_x(\mathcal{M}_x)^\perp$, $t \mapsto J(x+th)$ non-smooth at $t = 0$.

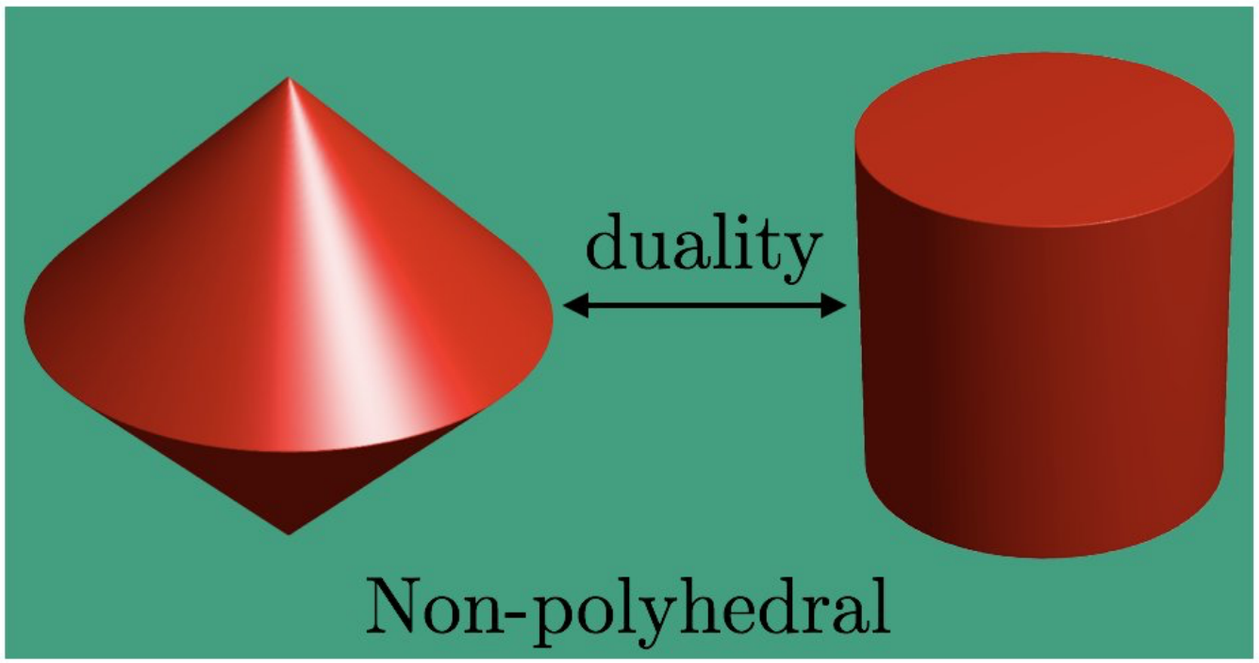
(iii) ∂J is continuous on $\mathcal{M}_x$ around x.

[Lewis 2003]

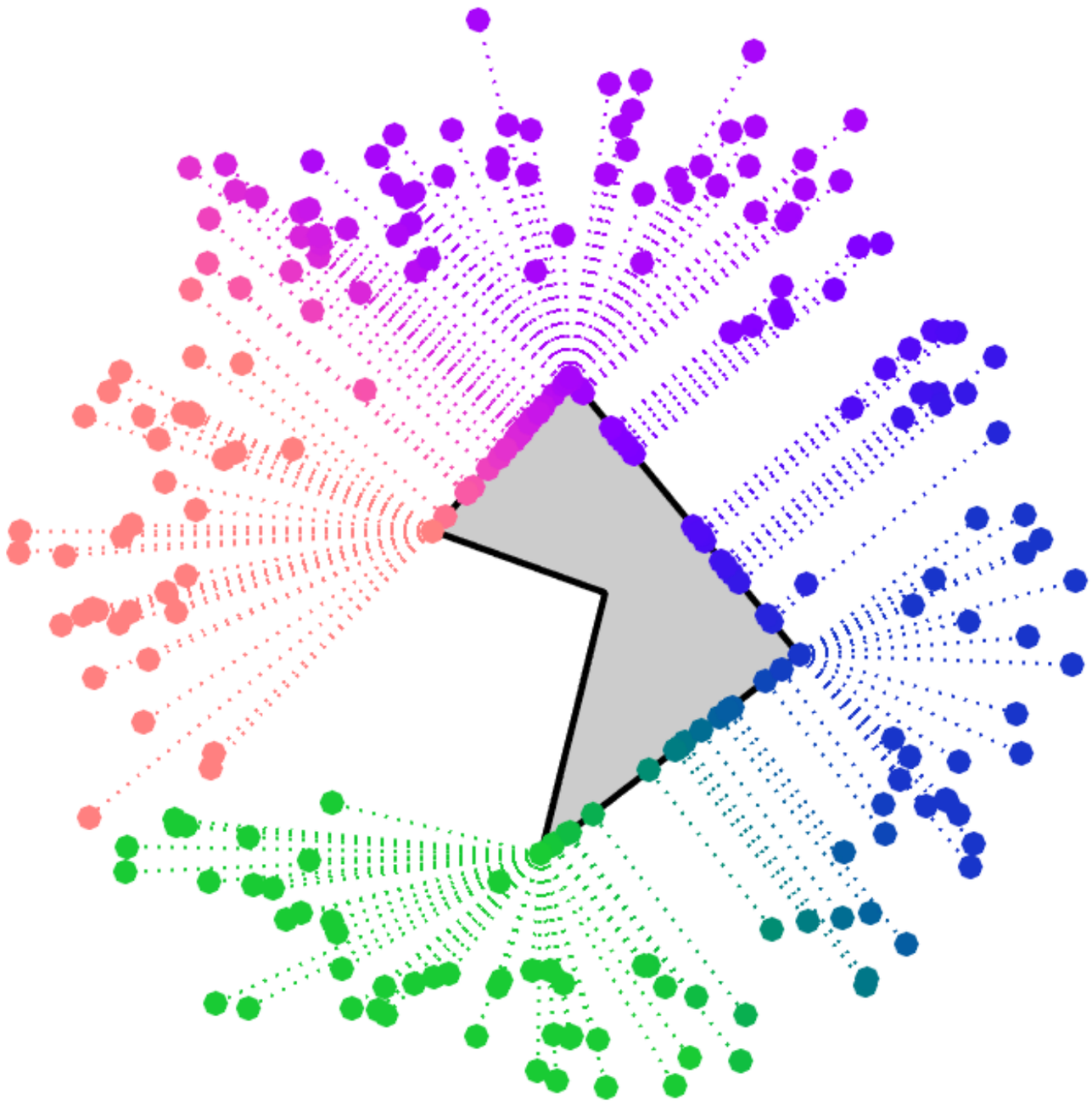


Partial smoothness [Lewis]: the correct notion of "piecewise smooth functions" in optimization. https://hal.archives-ouvertes.fr/hal-00987293



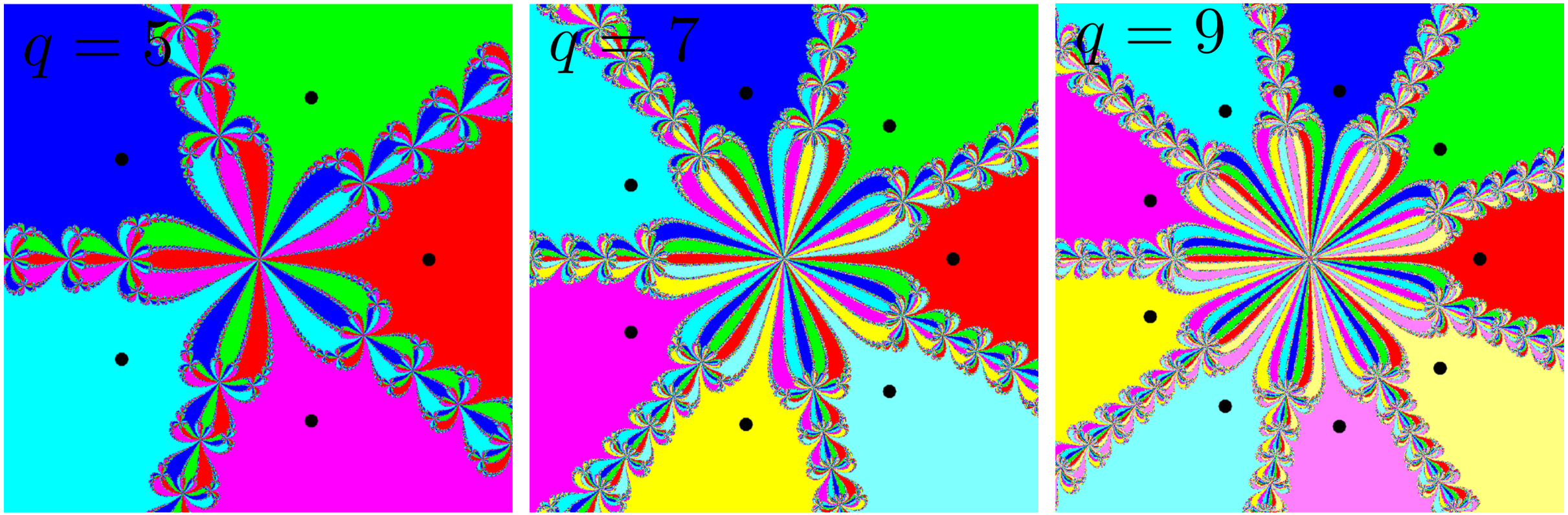


Mirror-stratifiability (Drusvyatskiy/Lewis): generalizes duality of polyhedra. Should become mainstream! https://arxiv.org/abs/1707.03194



The unreasonable effectiveness of non-smooth optimization: sharp singularities attract solutions of optimization problems.

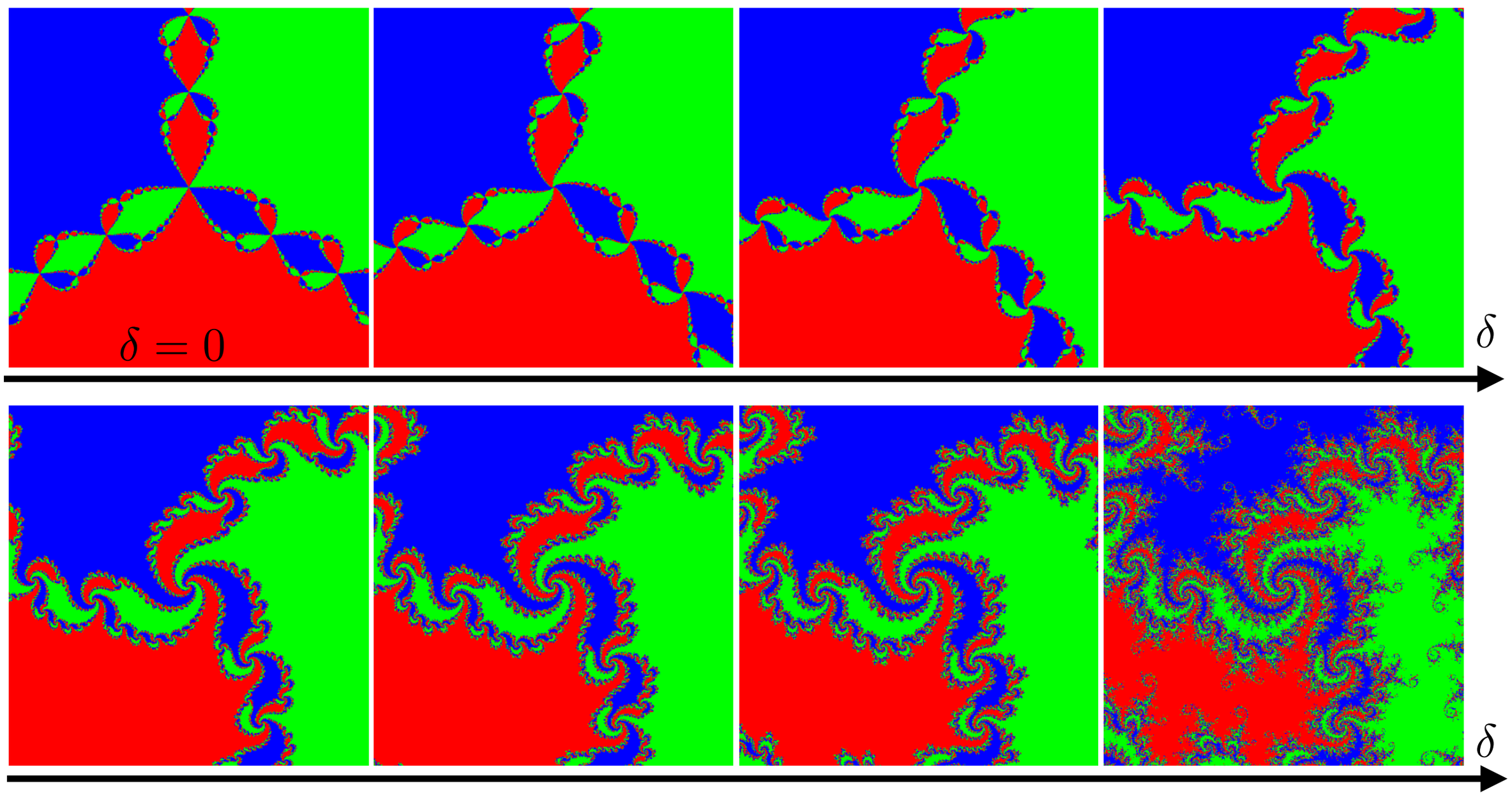
Newton method: $z_{k+1} = z_k - \frac{f(z_k)}{f'(z_k)}$



Attraction bassins for $f(z) = z^q - 1$

Understanding global convergence of Newton method is hard ... Code provided. https://goo.gl/xbpJCe

"Twisted" Newton: $z_{k+1} = z_k - (1 + \delta e^{\mathrm{i}\theta}) \frac{f(z_k)}{f'(z_k)}$ $\qquad f(z) = z^3 - 1$



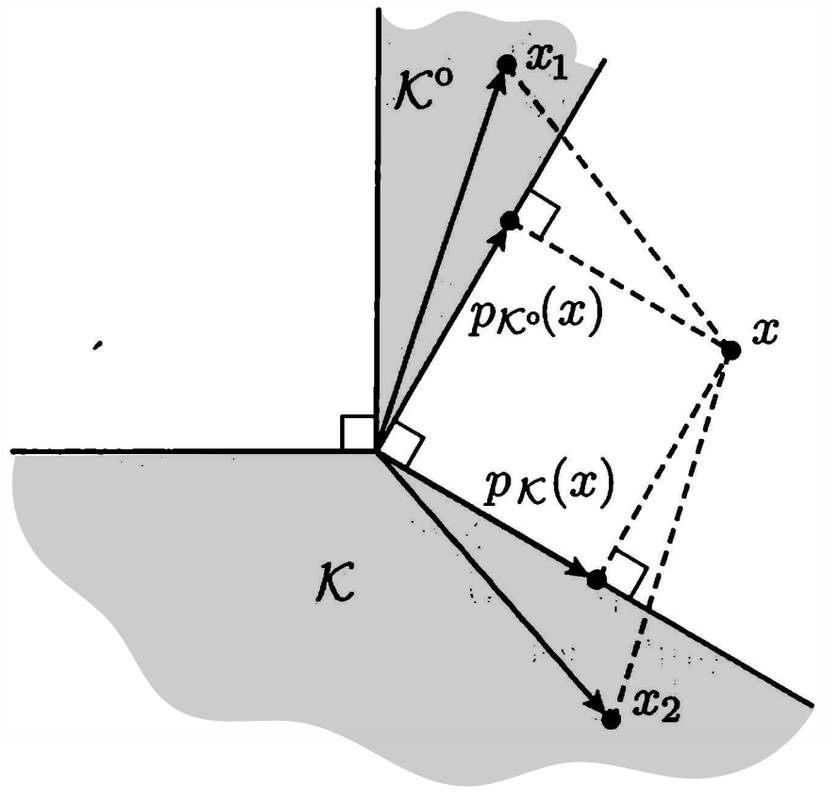
"Twisted" Newton iterations generate nice fractal patterns! Colors display bassin of attraction of cubic roots of unity.

Polar cone:

$$\mathcal{K}^\circ \stackrel{\text{def.}}{=} \{x \;;\; \forall y \in \mathcal{K}, \langle x, y\rangle \leqslant 0\}$$

Theorem: (Moreau's decomposition)

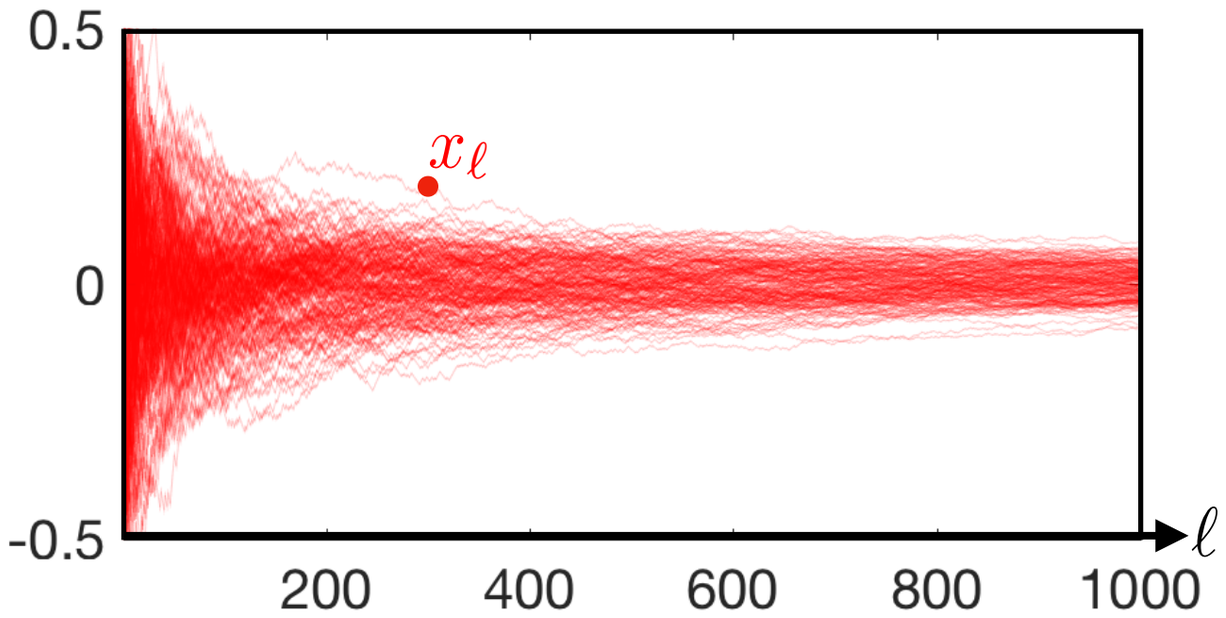
$$x = \mathrm{Proj}_{\mathcal{K}}(x) +^{\perp} \mathrm{Proj}_{\mathcal{K}^\circ}(x)$$

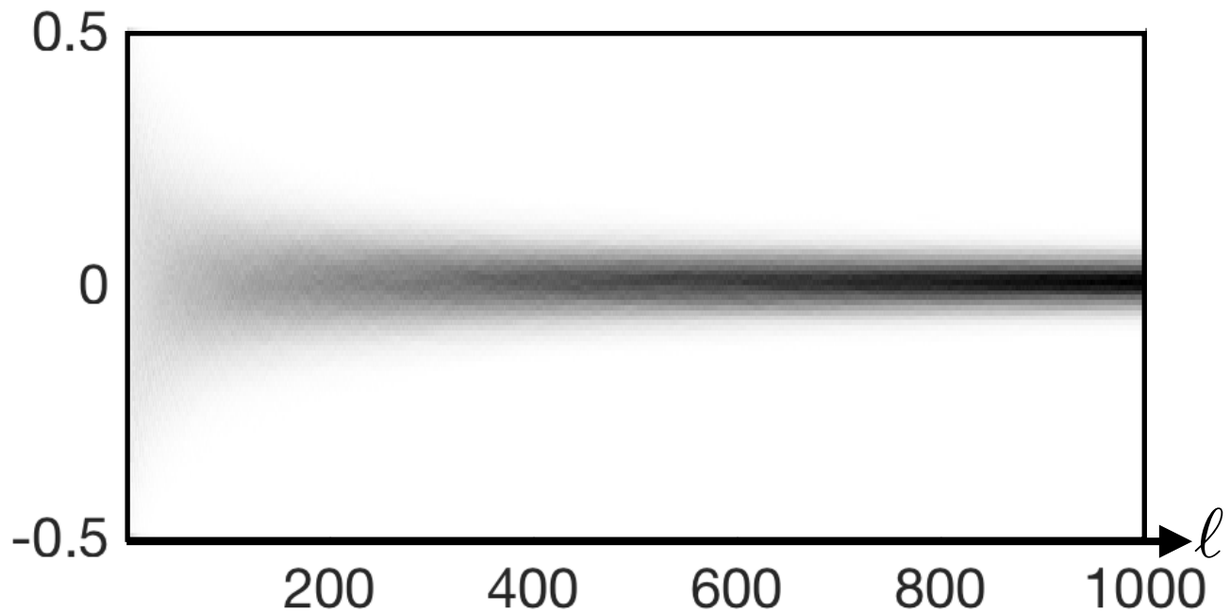


Moreau's decomposition generalizes orthogonal decomposition from linear spaces to convex cones.

$$\min_{x\in\mathbb{R}} \underbrace{(x+1)^2}_{=f_1(x)} + \underbrace{(x-1)^2}_{=f_2(x)}$$

$$x_{\ell+1} \overset{\text{def.}}{=} \begin{cases} x_\ell - \frac{1}{\ell}\nabla f_1(x_\ell) \text{ with proba } \frac{1}{2} \\ x_\ell - \frac{1}{\ell}\nabla f_2(x_\ell) \text{ with proba } \frac{1}{2} \end{cases}$$





Stochastic gradient descent dynamic on a simple 1-D example.

Gradient descent dynamic:

$$x(t) = -\nabla f(x(t)))$$

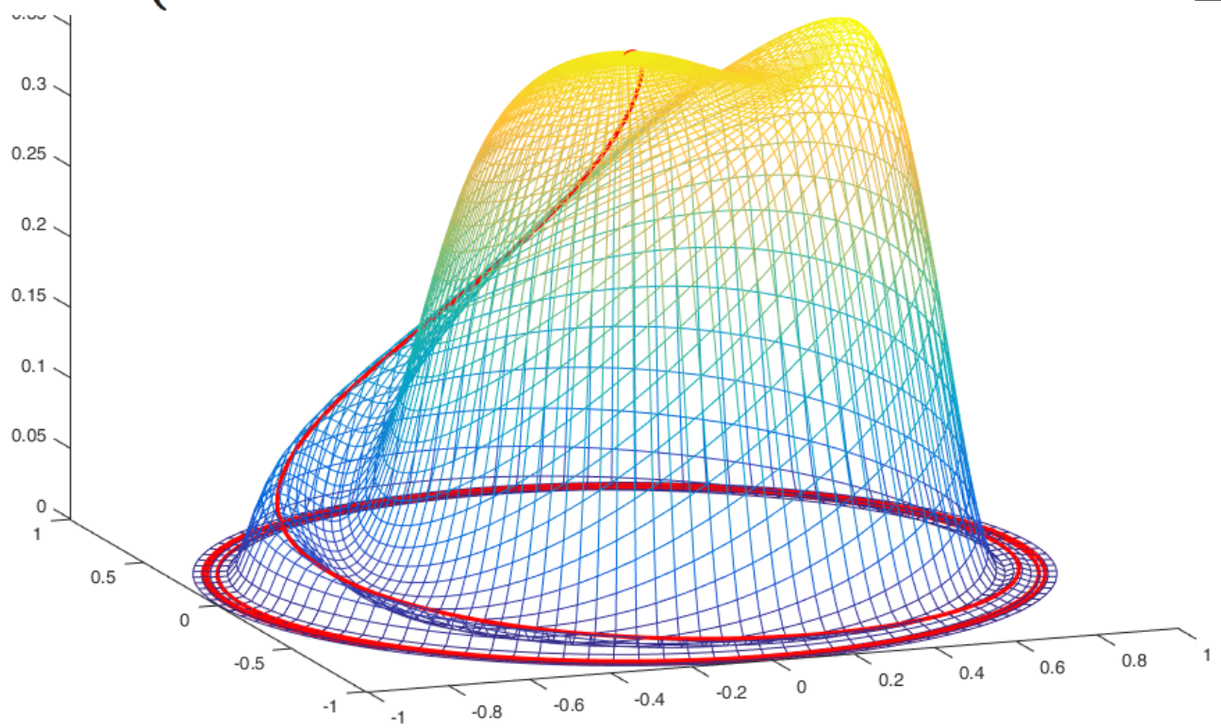
On the "mexican hat":

$$x(t) = r(t)(\cos(\theta(t)), \sin(\theta(t))$$

$$\theta(t) = (1 - r(t))^{-2}$$

$\rightarrow \text{Length}(x) = +\infty$

$$f(r,\theta) := \begin{cases} e^{-\frac{1}{1-r^2}}\left[1 - \frac{4r^4}{4r^4+(1-r^2)^4}\sin\left(\theta - \frac{1}{1-r^2}\right)\right] & \text{if } r < 1, \\ 0 & \text{if } r \geq 1, \end{cases}$$



Gradient descent does not always converge. Trajectories can have infinite length. Picture & insights from @GuillaumeG_ http://epubs.siam.org/doi/abs/10.1137/040605266

$$f : z \stackrel{\text{def.}}{=} x + \mathrm{i}y \in \mathbb{C} \longmapsto f(z) \in \mathbb{C}$$

$$\frac{\partial}{\partial z} \stackrel{\text{def.}}{=} \frac{1}{2}\left(\frac{\partial}{\partial x} - \mathrm{i}\frac{\partial}{\partial y}\right) \qquad \frac{\partial}{\partial \bar{z}} \stackrel{\text{def.}}{=} \frac{1}{2}\left(\frac{\partial}{\partial x} + \mathrm{i}\frac{\partial}{\partial y}\right)$$

Proposition: f holomorphic $\Leftrightarrow \frac{\partial f}{\partial \bar{z}} = 0.$

Examples:

$$\frac{\partial z}{\partial z} = 1 \quad \frac{\partial \bar{z}}{\partial z} = 0 \quad \frac{\partial z}{\partial \bar{z}} = 0 \quad \frac{\partial \bar{z}}{\partial \bar{z}} = 1$$

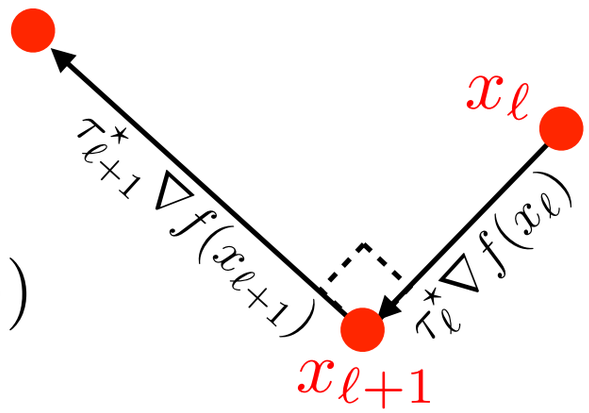
$$\frac{\partial cz}{\partial z} = c \qquad \frac{\partial |z|^2}{\partial z} = \frac{\partial (z\bar{z})}{\partial z} = \bar{z}$$

Wirtinger derivatives: treat complex valued functions of complex variables as if they were real functions. Very convenient.

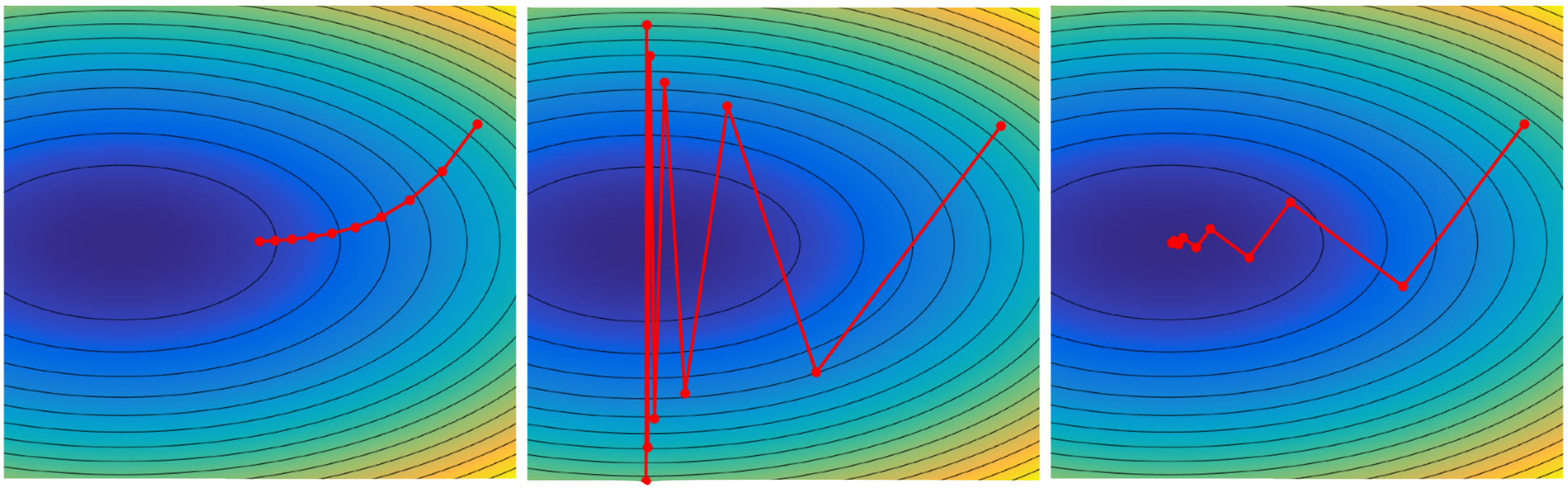
https://en.wikipedia.org/wiki/Wirtinger_derivatives

$$x_{\ell+1} = x_\ell - \tau_\ell \nabla f(x_\ell)$$

$$\tau_\ell^\star = \operatorname*{argmin}_\tau f(x_\ell - \tau \nabla f(x_\ell))$$



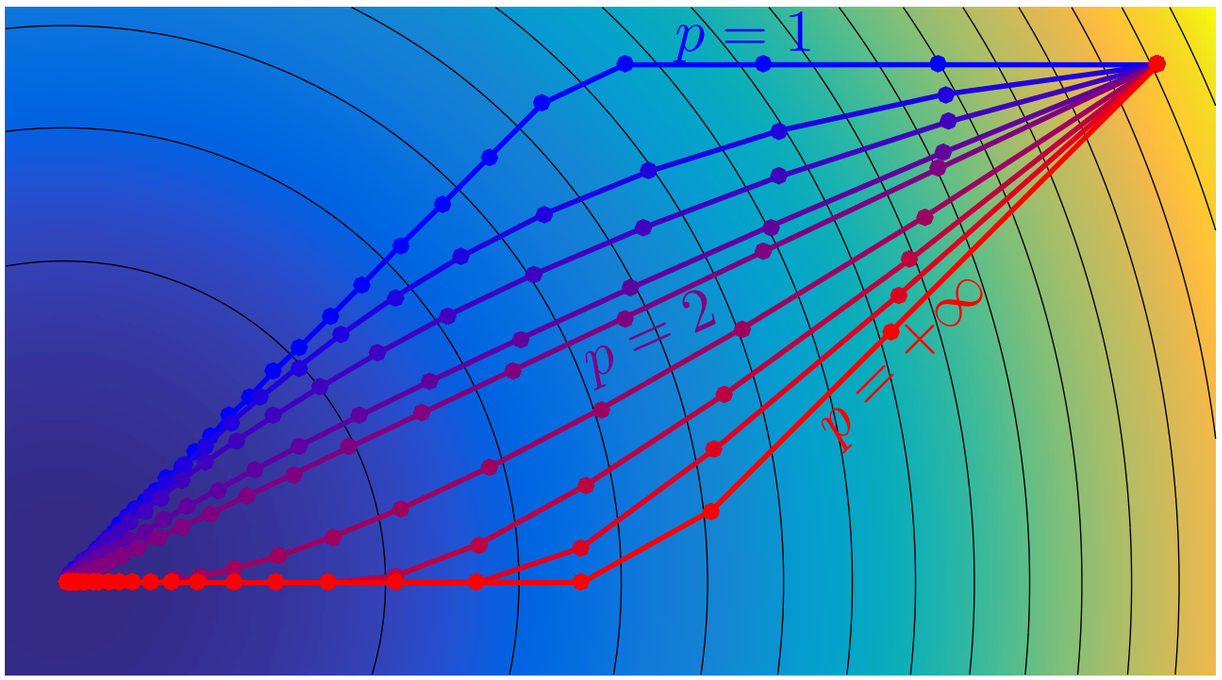
$$\nabla f(x_\ell) \perp \nabla f(x_{\ell+1})$$



Small τ_ℓ | Large τ_ℓ | Optimal $\tau_\ell = \tau_\ell^\star$

Step size selection is important for gradient descent on ill-conditioned functions.

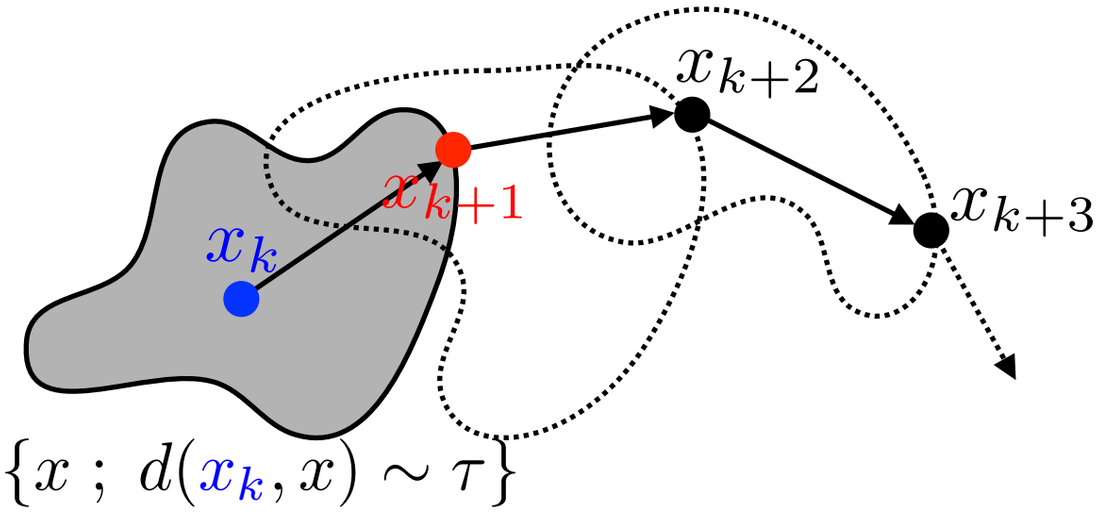
Metric space $(\mathcal{X}, d)$, minimize $F(x)$ on $\mathcal{X}$.



$F(x) = \|x\|^2$ on $(\mathcal{X} = \mathbb{R}^2, \|\cdot\|_p)$

Implicit Euler step:

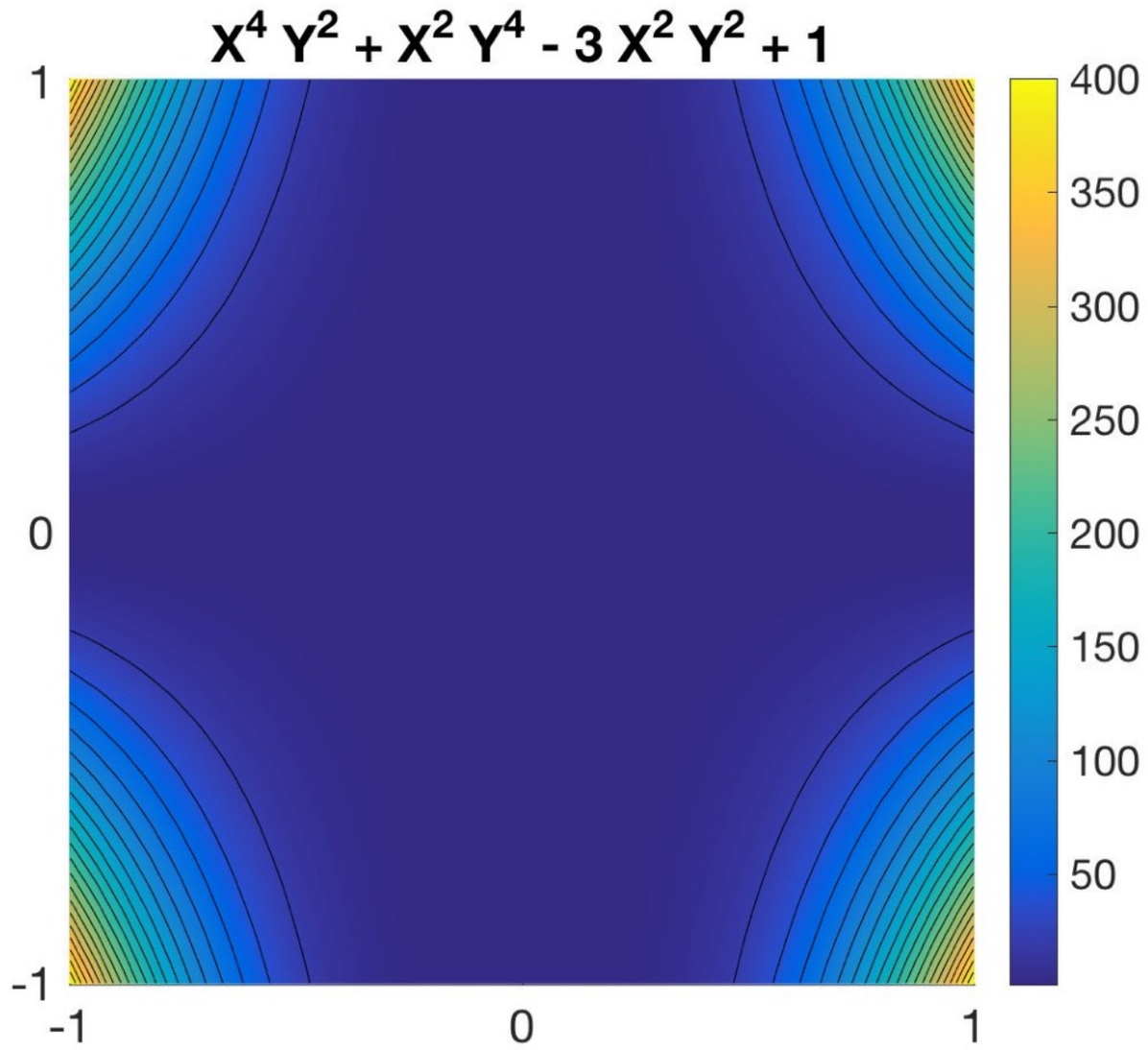
$$x_{k+1} \overset{\text{def.}}{=} \operatorname*{argmin}_{x \in \mathcal{X}} d(x_k, x)^2 + \tau F(x)$$



Gradient flows on metric spaces: define descent methods on non-Euclidean domains without using a gradient. A key tool studied in details by Ambrosio, Gigli and Savare.

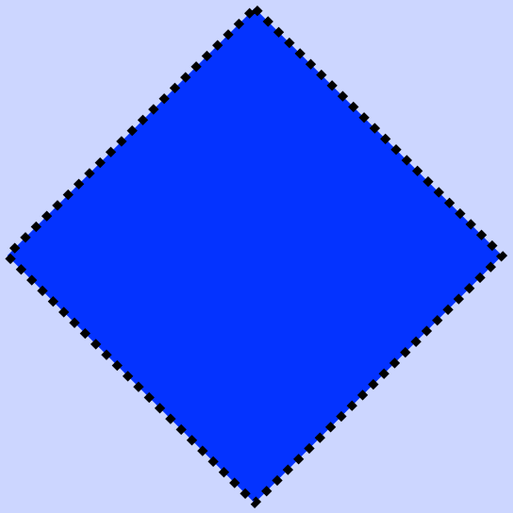
http://www.springer.com/la/book/9783764387211

Matrix and Polynomials



Motzkin's polynomial is positive but is not a sum of squares.
#LifeIsHarderIn2D https://en.wikipedia.org/wiki/Positive_polynomial …

Vectors



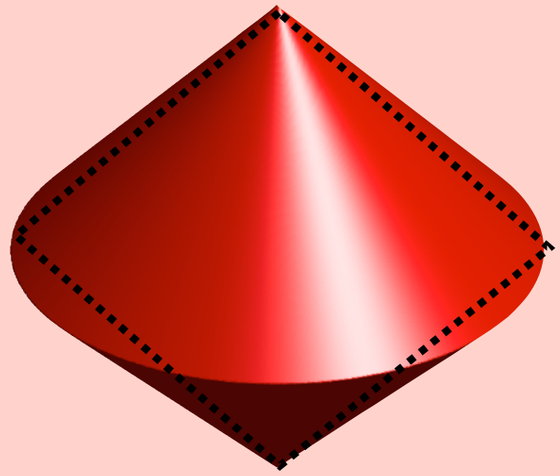
$$\|x\|_1 \leqslant 1$$

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

spectral

lift

Matrices



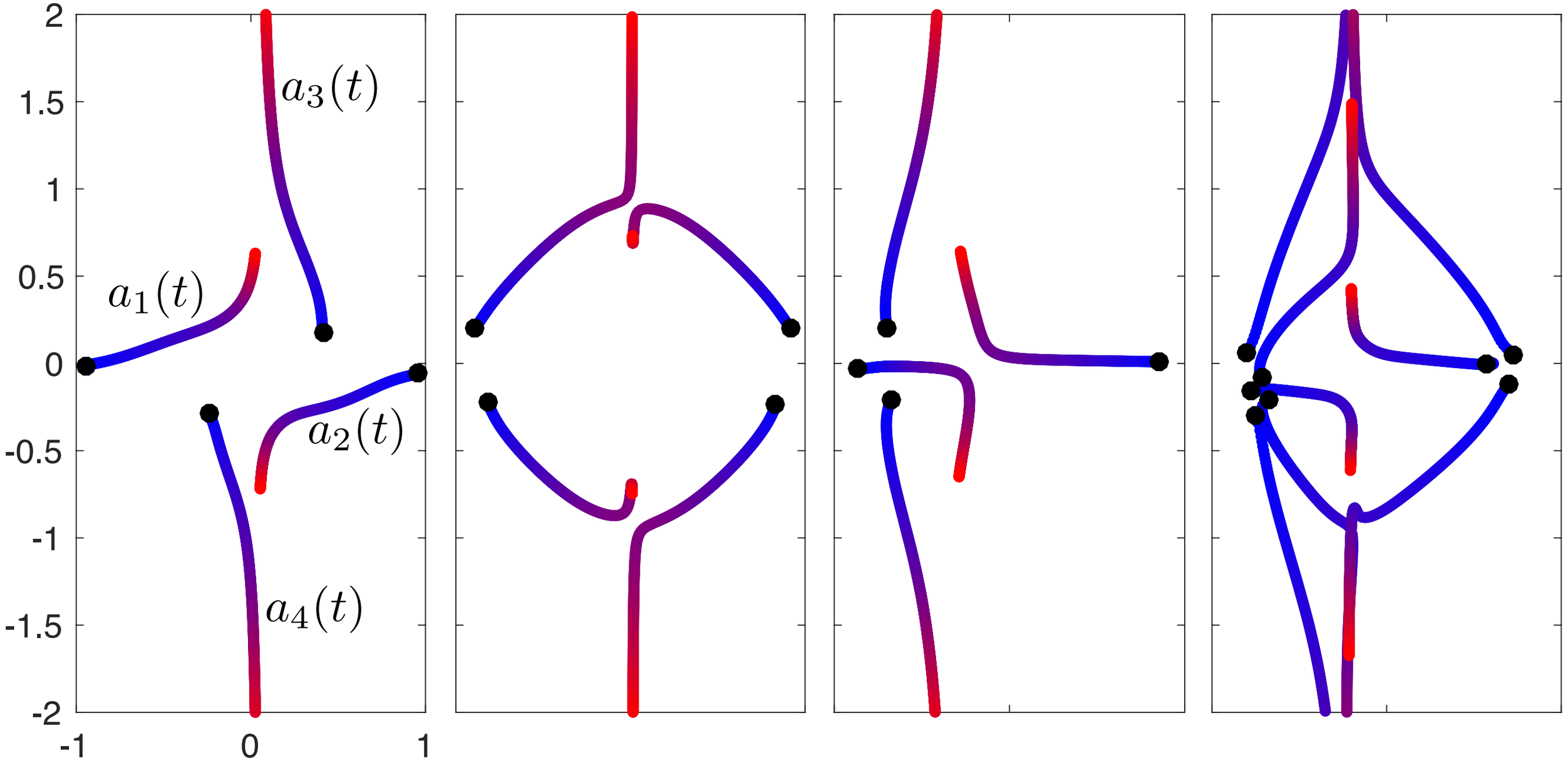
$$\|\mathrm{eig}(X)\|_1 \leqslant 1$$

$$X = \begin{pmatrix} X_1 & X_2 \\ X_2 & X_3 \end{pmatrix}$$

Reasonable properties (convexity, differential, stratifiability, partial-smooth...) lift from vectors to matrices! https://people.orie.cornell.edu/aslewis/publications/03-mathematics.pdf ...

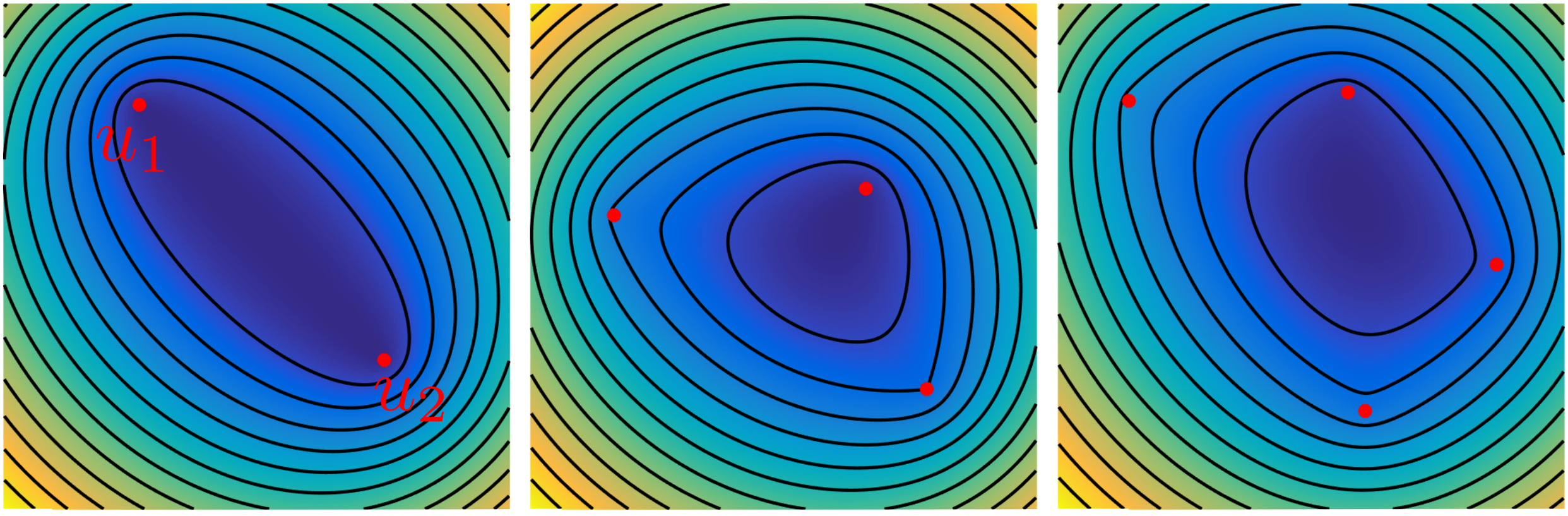
The QR algorithm is a gem from numerical matrix analysis. Showing convergence toward triangular matrix. https://en.wikipedia.org/wiki/QR_algorithm …

$$\partial_t P_t(x) = \partial_x^2 P_t(x) \qquad P_t(x) = \prod_i (x - a_i(t))$$



Heat eq on polynomials generates nice roots dynamics. Post by Terry Tao. Code provided!
https://terrytao.wordpress.com/2017/10/17/heat-flow-and-zeroes-of-polynomials/ …
https://goo.gl/hPE5uo

$$\left\{ (x,y) \in \mathbb{R}^2 \; ; \; \sum_{i=1}^{n} \sqrt{(x-u_k)^2 + (y-v_k)^2} \leqslant t \right\}$$

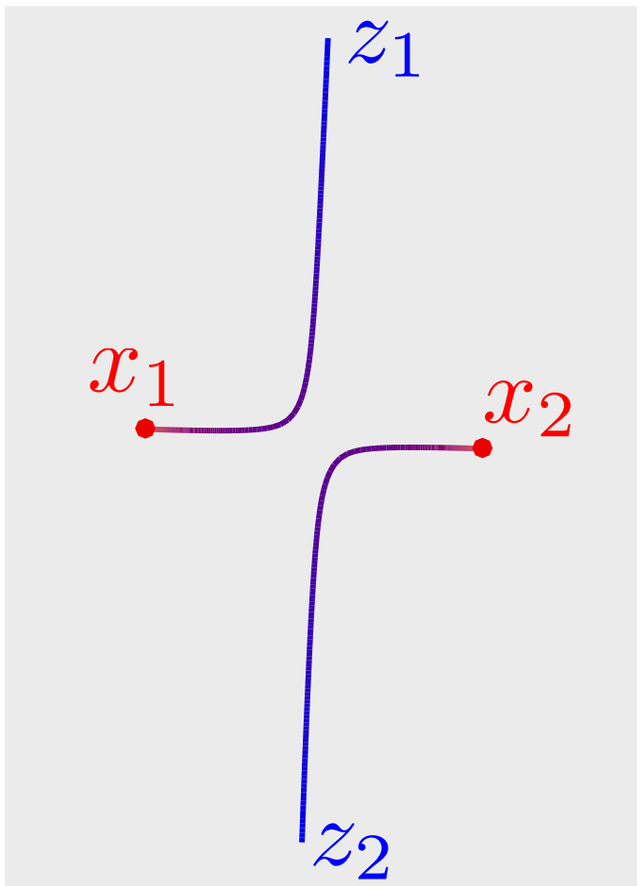


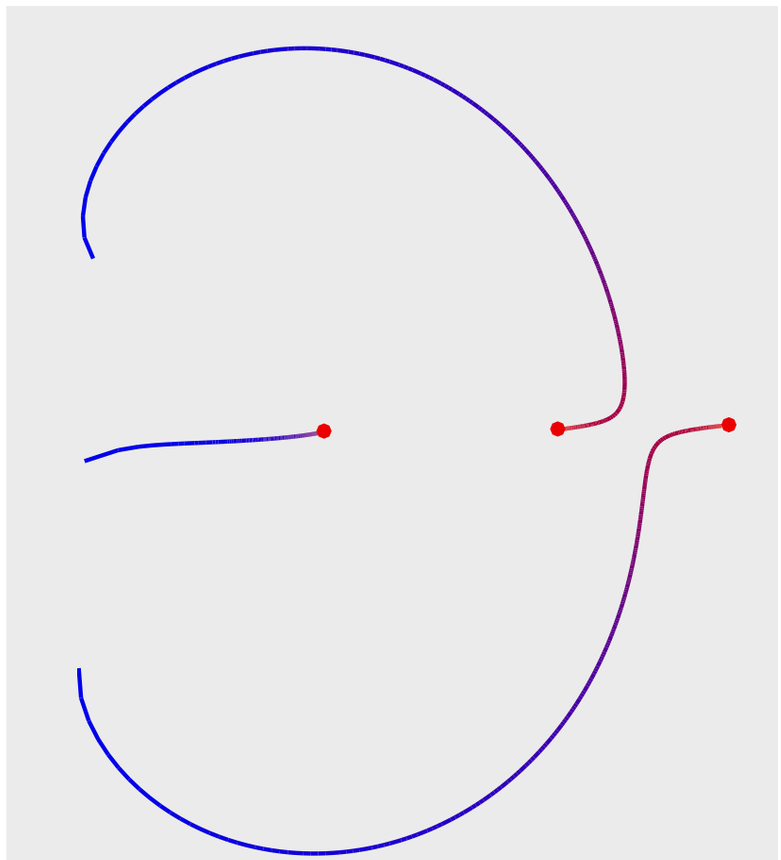
n-ellipses generalize ellipses with n foci. Spectrahedra (hence convex) algebraic curves of degree less than 2^n. https://en.wikipedia.org/wiki/N-ellipse

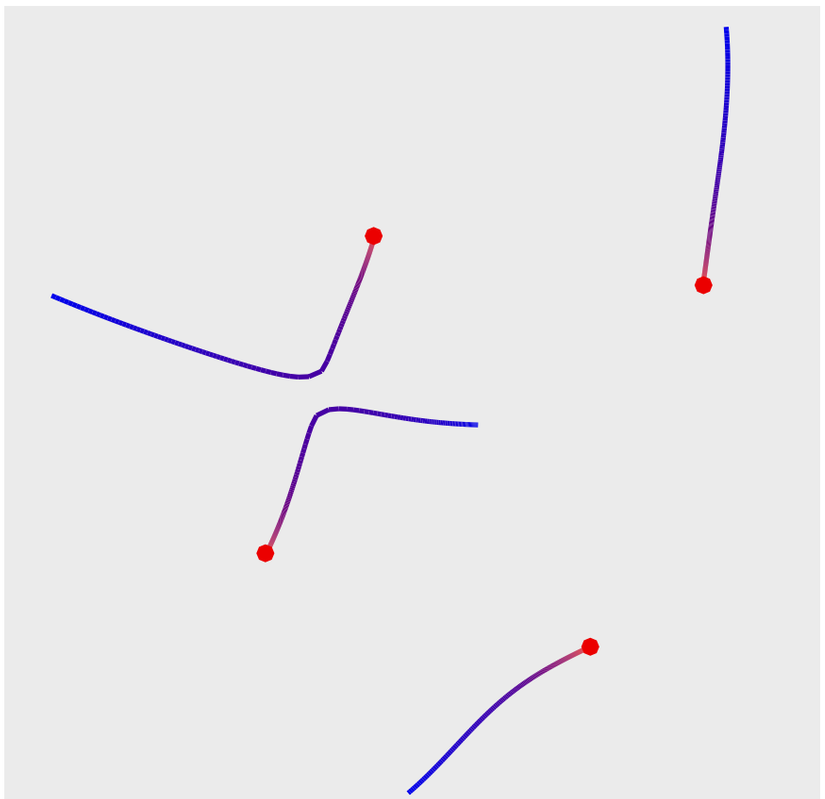
$$z_i \longmapsto z_i - \frac{P(z_i)}{\prod_{j \neq i}(z_i - z_j)}$$

Theorem:

$$(z_i)_i \longrightarrow \mathrm{Roots}(P)$$







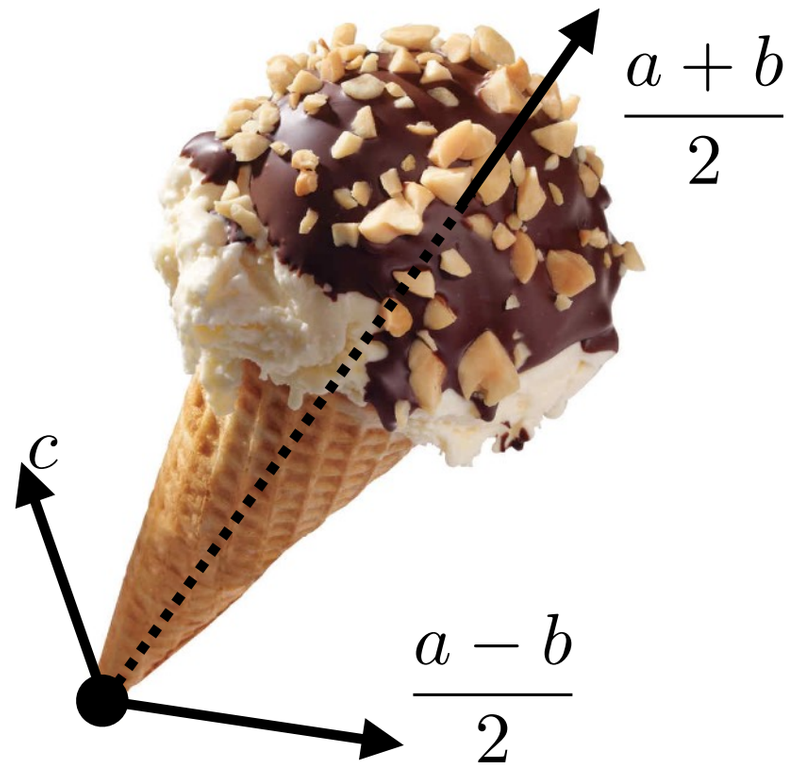
$$P(x) = \prod_i (x - x_i)$$

Durand–Kerner method: surprisingly simple way of computing all the roots of a polynomial. https://en.wikipedia.org/wiki/Durand%E2%80%93Kerner_method …

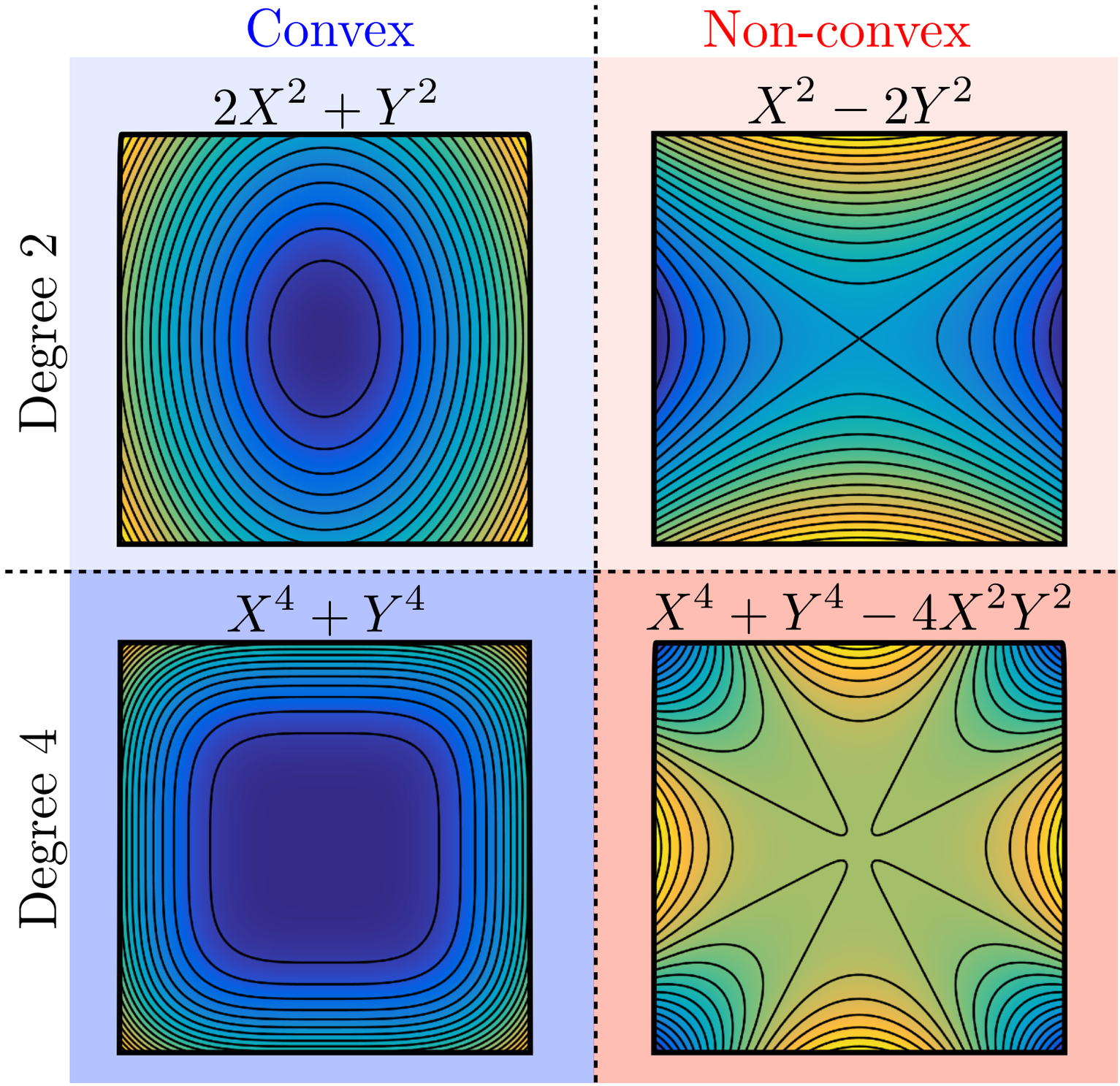
$$\left\{(a,b,c)\ ;\ \forall x, a+2cx+bx^2 \geqslant 0\right\}$$

$$= \left\{(a,b,c)\ ;\ \begin{pmatrix} a & c \\ c & b \end{pmatrix} \succeq 0\right\}$$

$$= \left\{(a,b,c)\ ;\ \left(\frac{a-b}{2}\right)^2 + c^2 \leqslant \left(\frac{a+b}{2}\right)^2\right\}$$



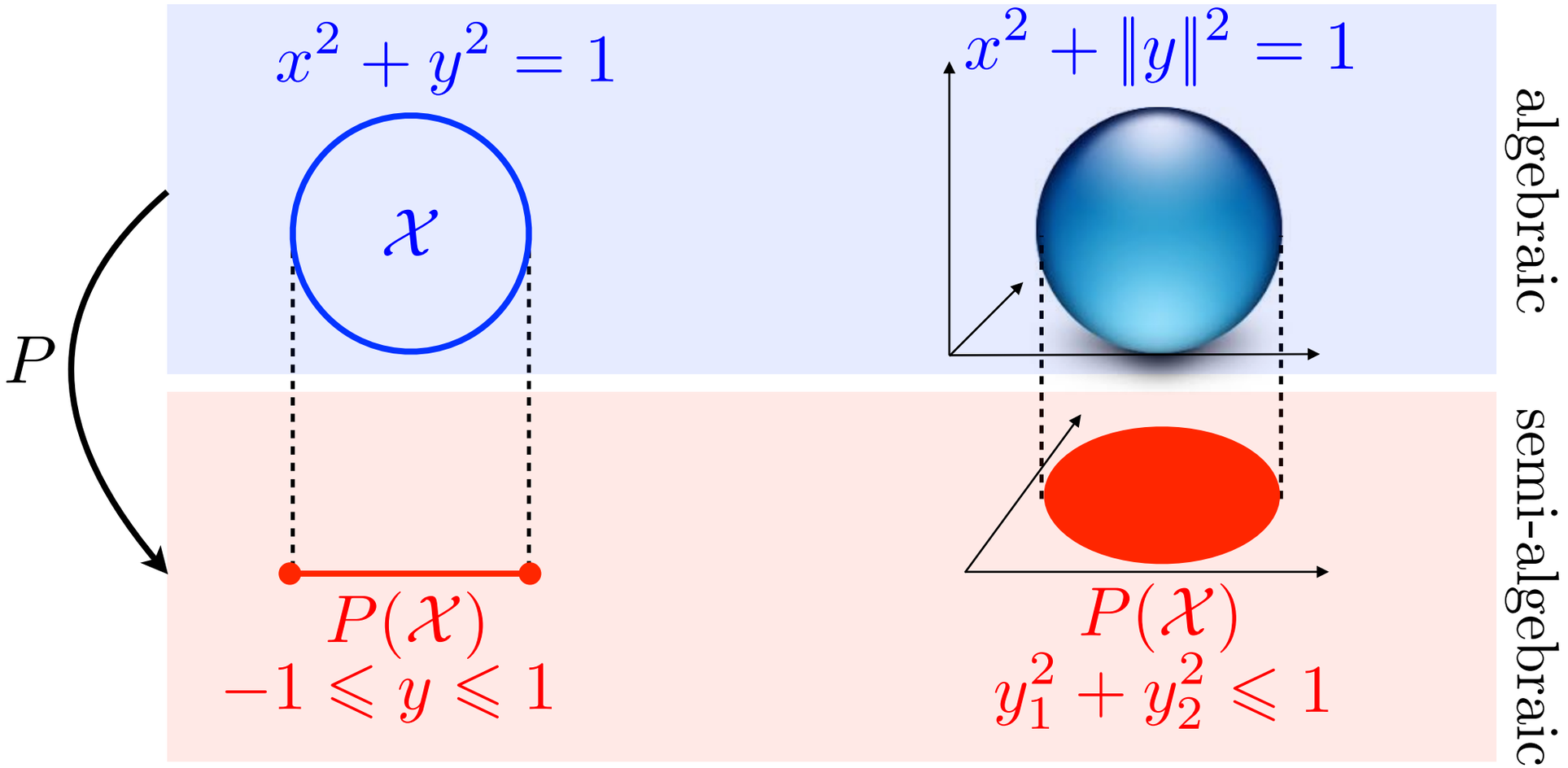
I knew it! Ice-creams, positive polynomials and PSD matrices all have the same flavor!



Checking convexity of degree 2 polynomials is trivial.
Checking convexity of degree 4 polynomials is NP-hard.
http://web.mit.edu/~a_a_a/Public/Publications/convexity_nphard.pdf

$$\mathcal{X} \subset \mathbb{R}^n \text{ semi-algebraic} \implies P(\mathcal{X}) \in \mathbb{R}^{n-1} \text{ semi-algebraic.}$$

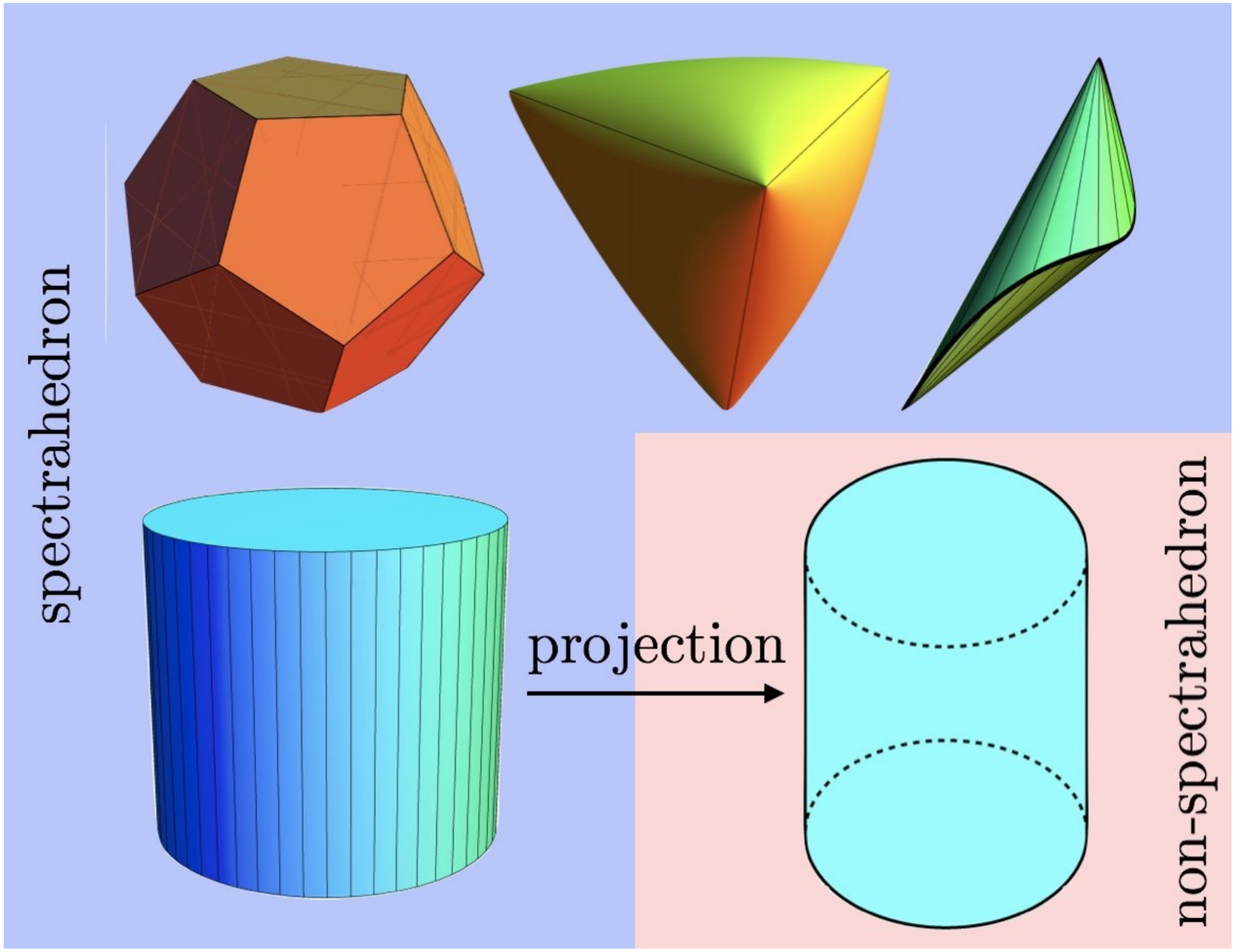
$$P : (x, y) \in \mathbb{R} \times \mathbb{R}^{n-1} \longmapsto y \in \mathbb{R}^{n-1}$$



Tarski–Seidenberg theorem: semi-algebraicity is stable by projection. The most fundamental result in semi-algebraic geometry. Equivalent to the elimination of the existential quantifier.

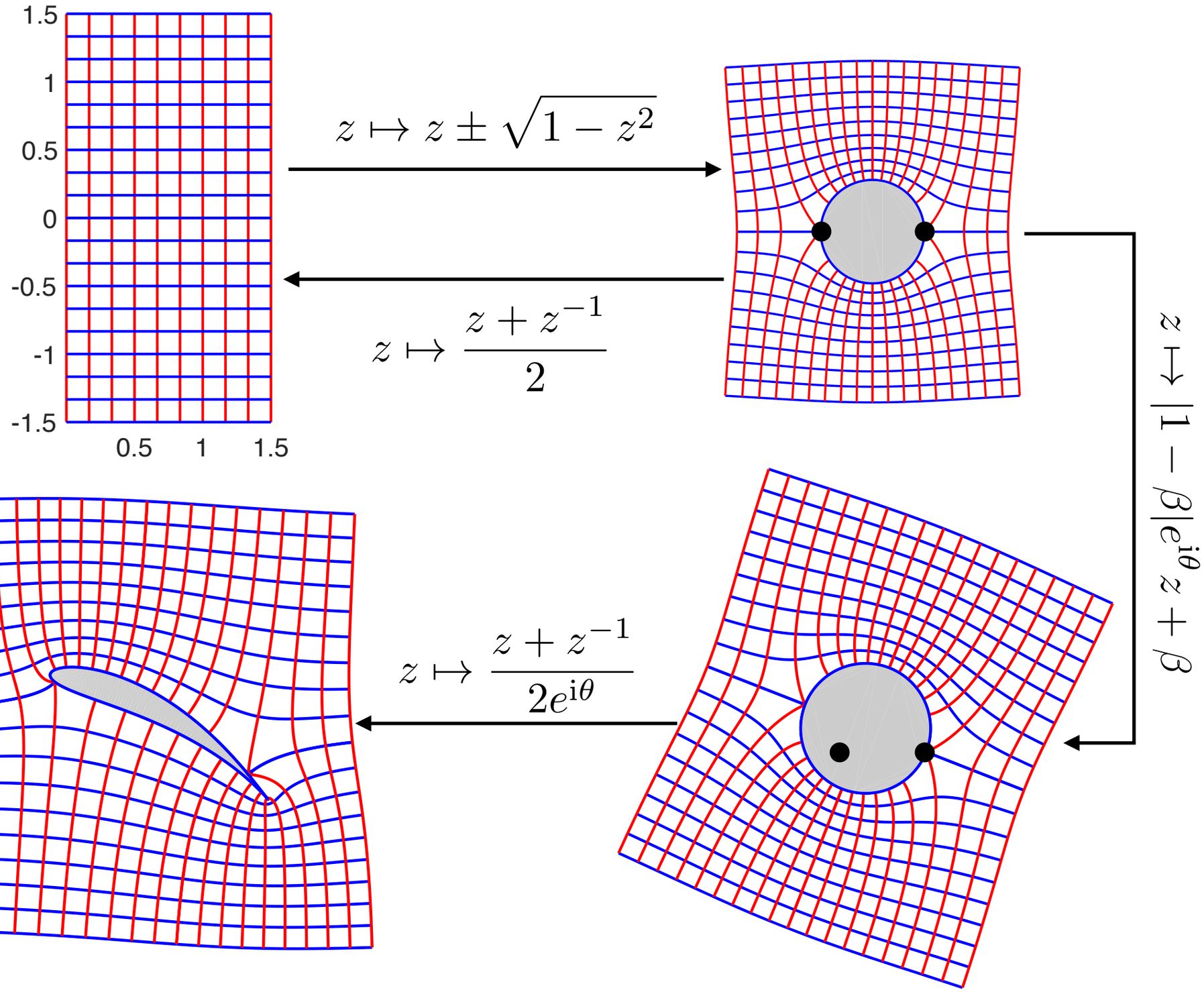
https://en.wikipedia.org/wiki/Tarski%E2%80%93Seidenberg_theorem

Geometry

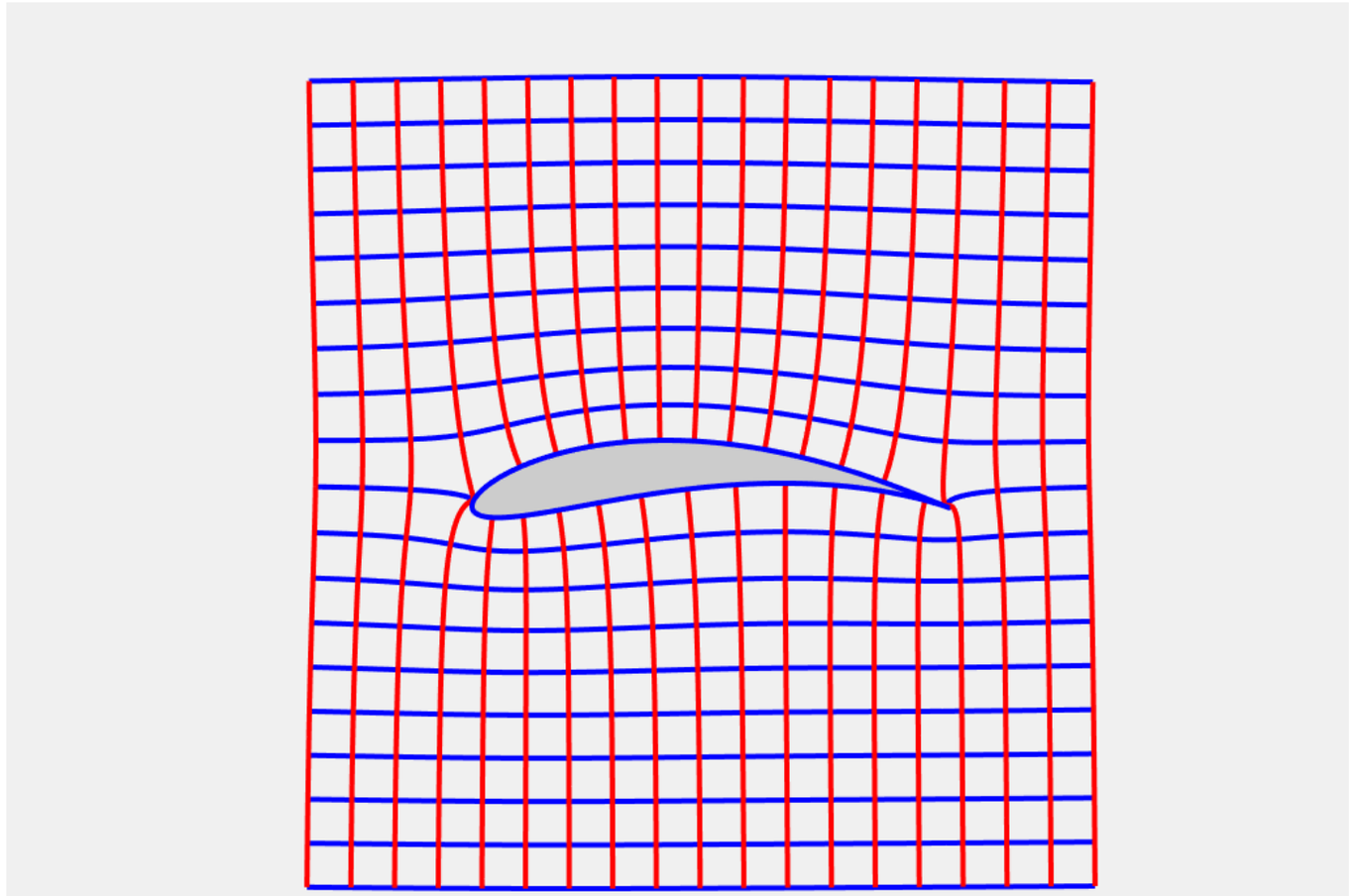


Spectrahedra are matrix generalizations of polyhedra. But the class of spectrahedra is not stable by projection. https://math.berkeley.edu/~bernd/WhatIsS

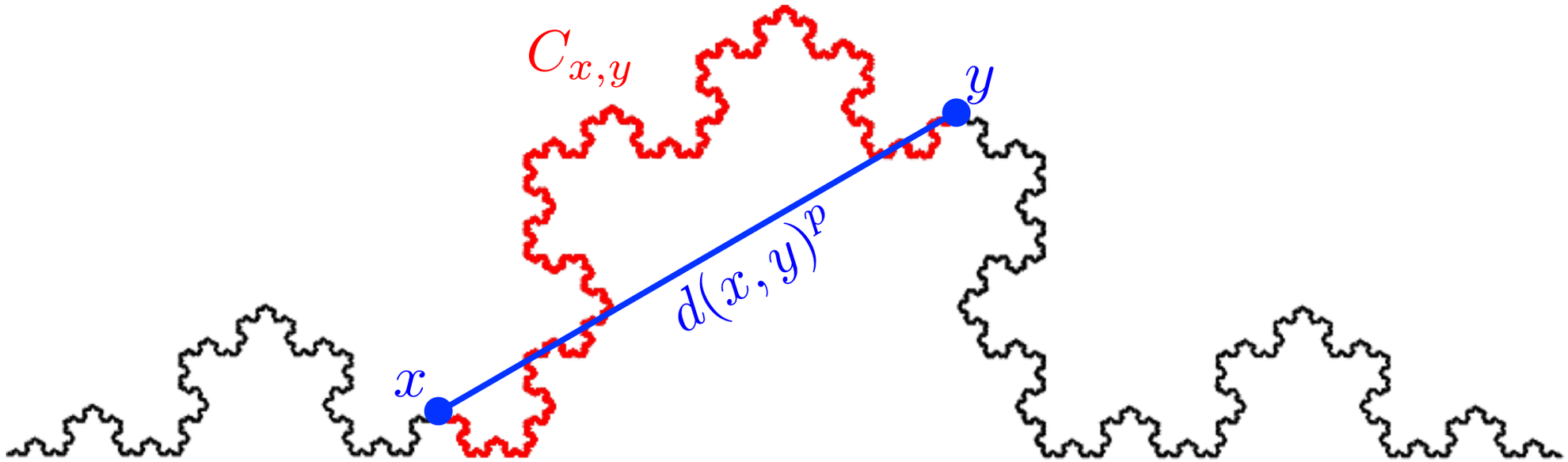
Joukowski Conformal Mapping



Joukowsky conformal map: solving airfoil design in the pre-computer era. https://en.wikipedia.org/wiki/Joukowsky_transform …

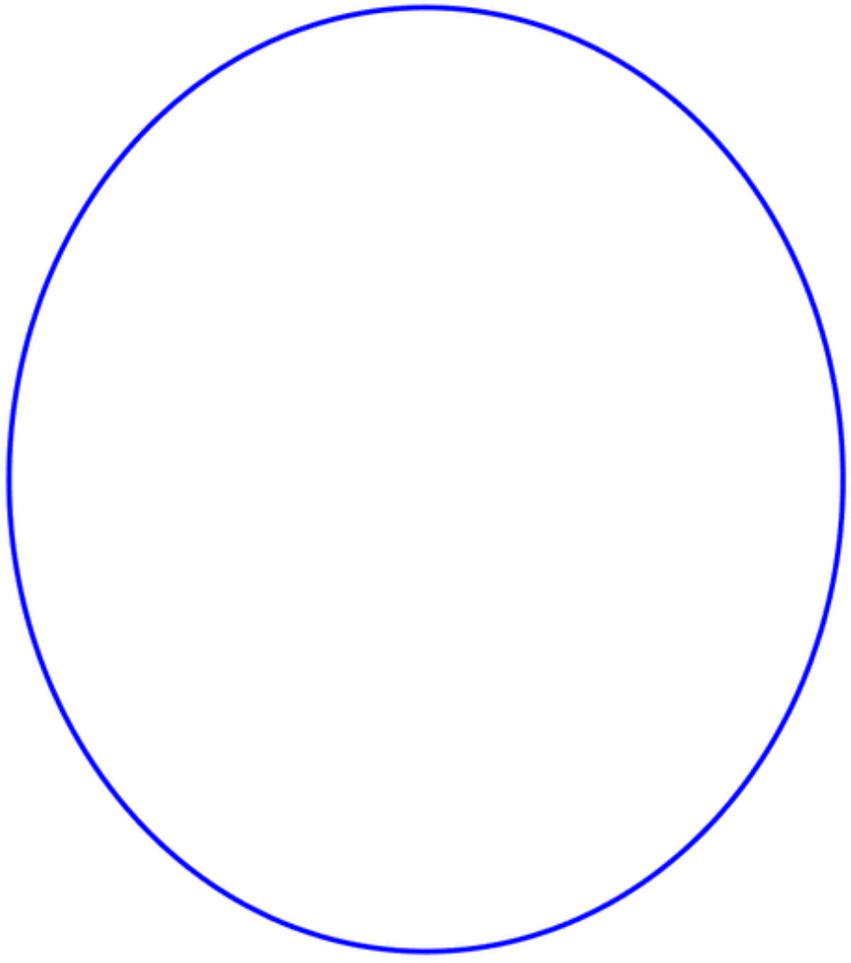


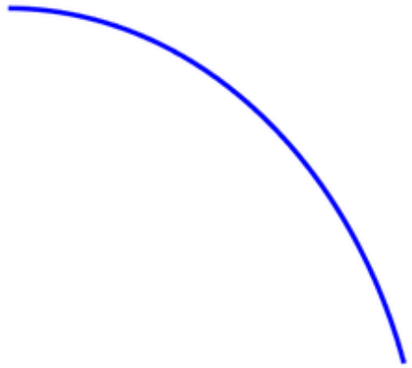
Rotating airfoil using Joukowsky conformal map. Code provided! https://goo.gl/W6igS5



$$d(x,y) = \text{Hausdorff}_p(C_{x,y})^{\frac{1}{p}} \qquad p = \frac{\log(4)}{\log(3)}$$

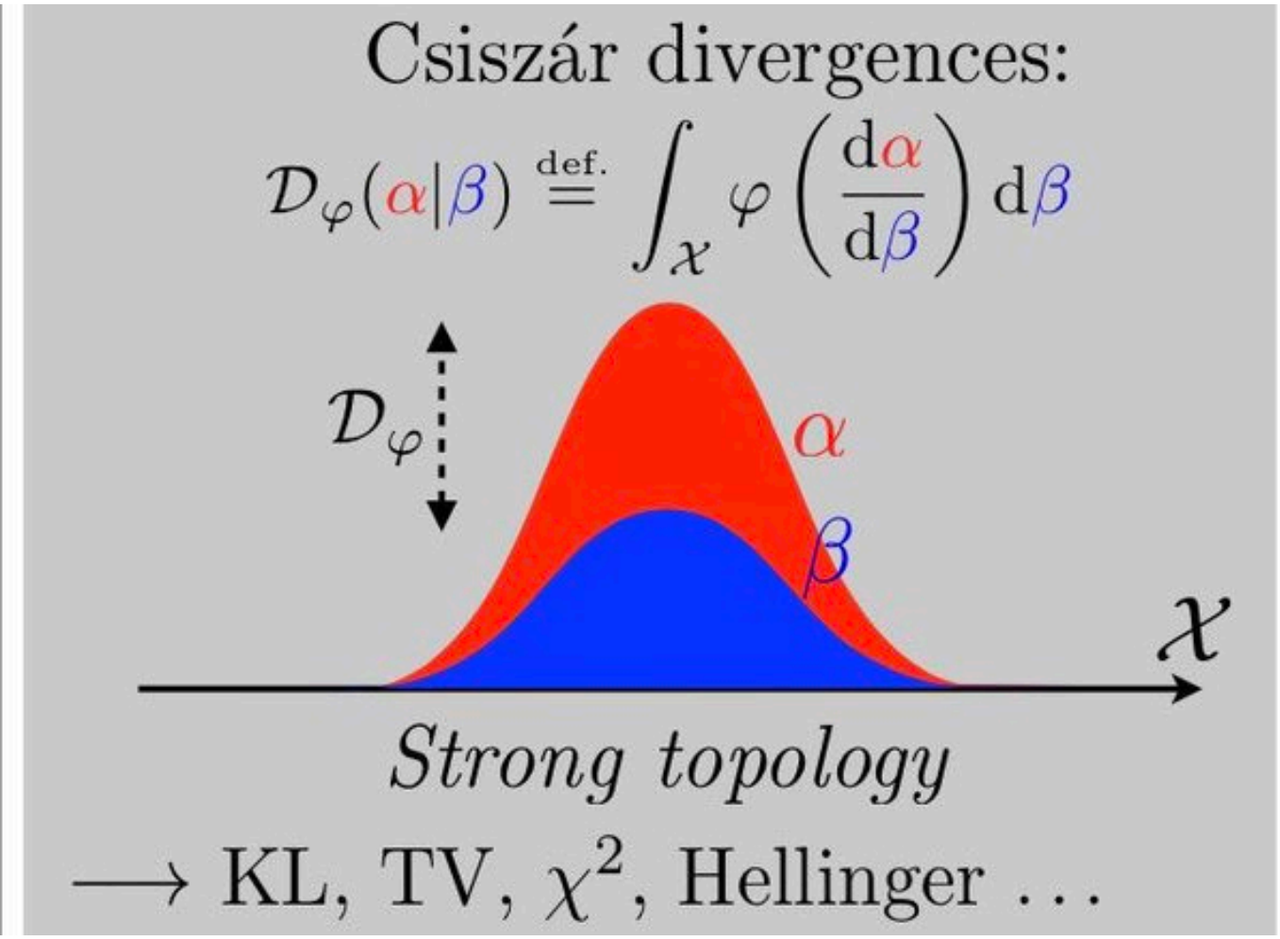
If d(x,y) is a distance, then d(x,y)^p for 0<p<1 is its snowflake distance. Naming bc of the von Koch fractal. https://www.emis.de/journals/AASF/Vol30/tyson.pdf …

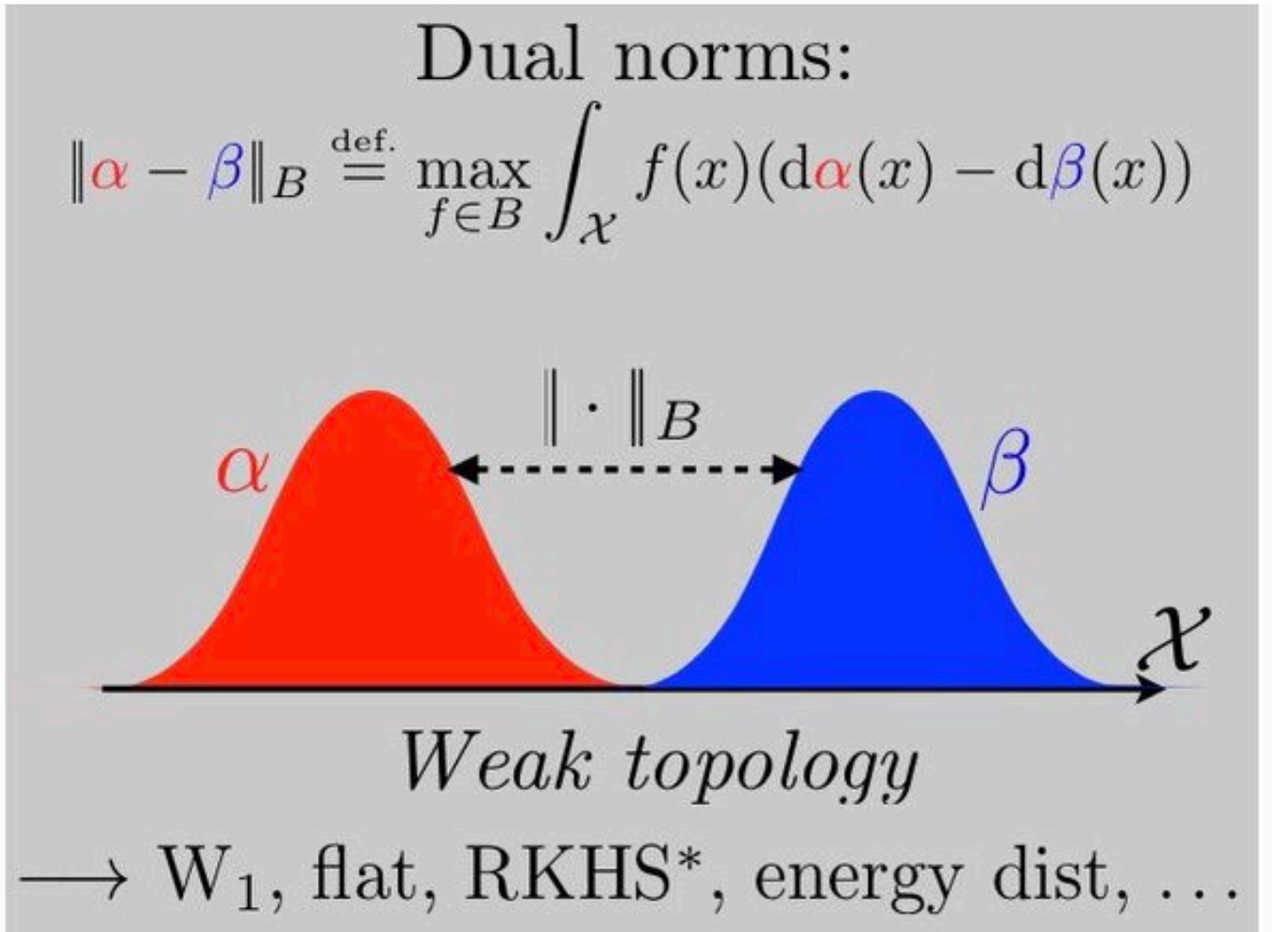




Spirograph (hypotrochoid) vs lissajou curves: messing around with a bunch of cos and sin.
https://en.wikipedia.org/wiki/Lissajous_curve
https://en.wikipedia.org/wiki/Hypotrochoid

Measures and Probability

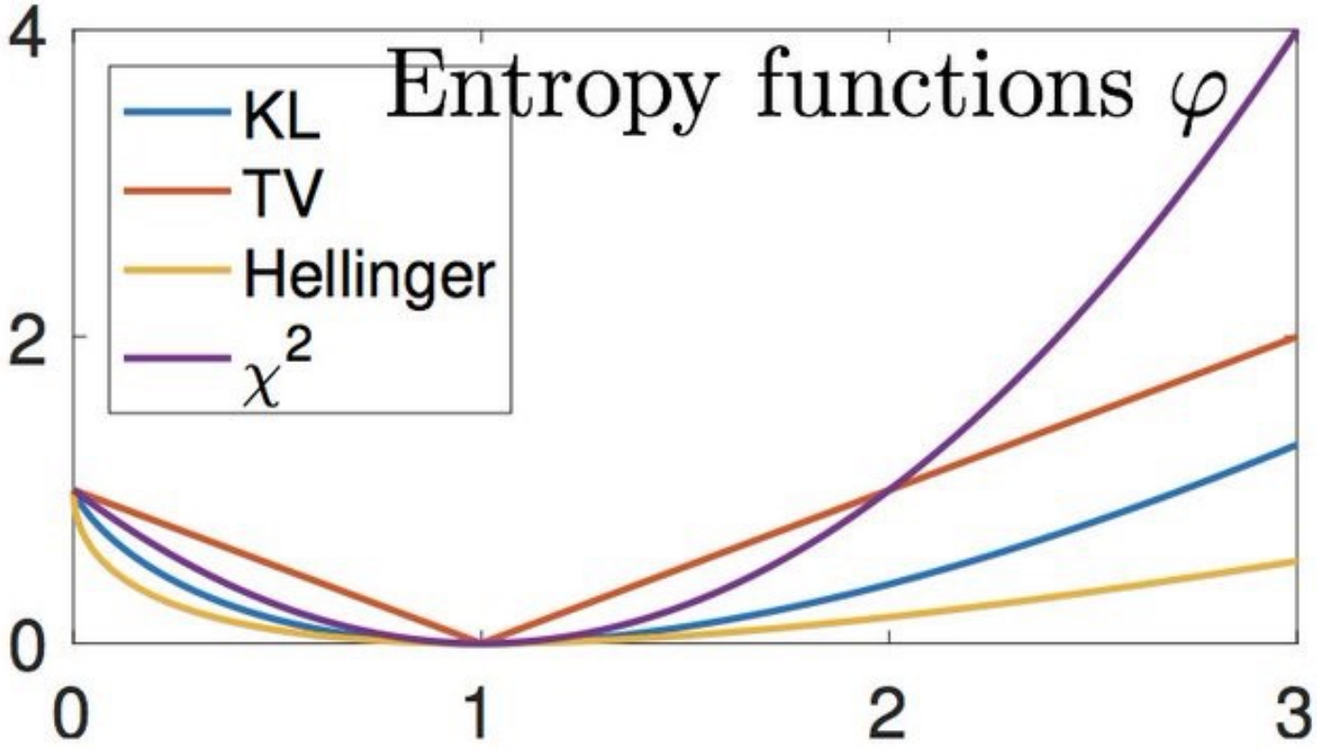


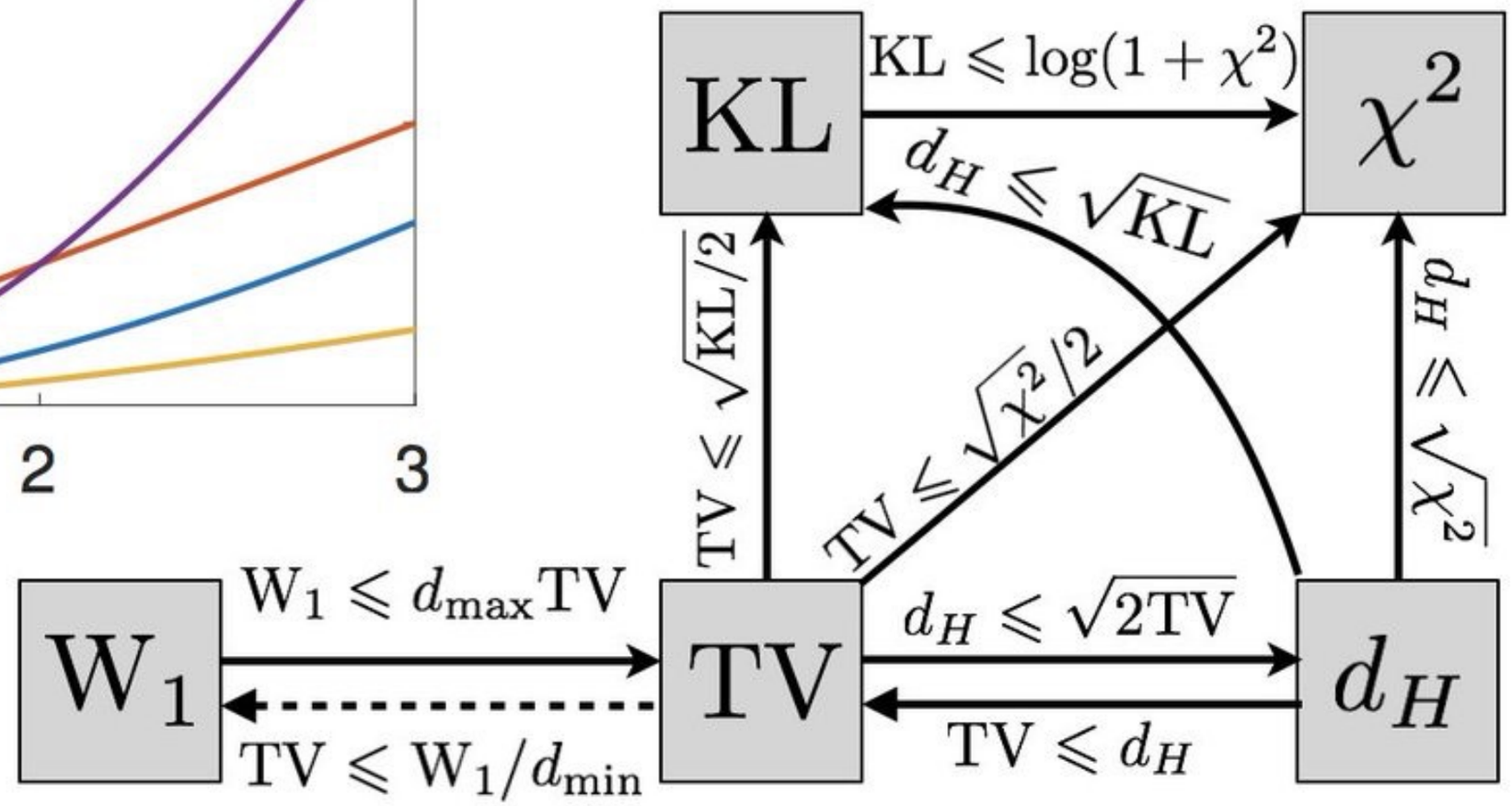


Comparing probability measures: vertical vs horizontal displacent / Csiszár divergence vs dual norm / strong vs weak topology

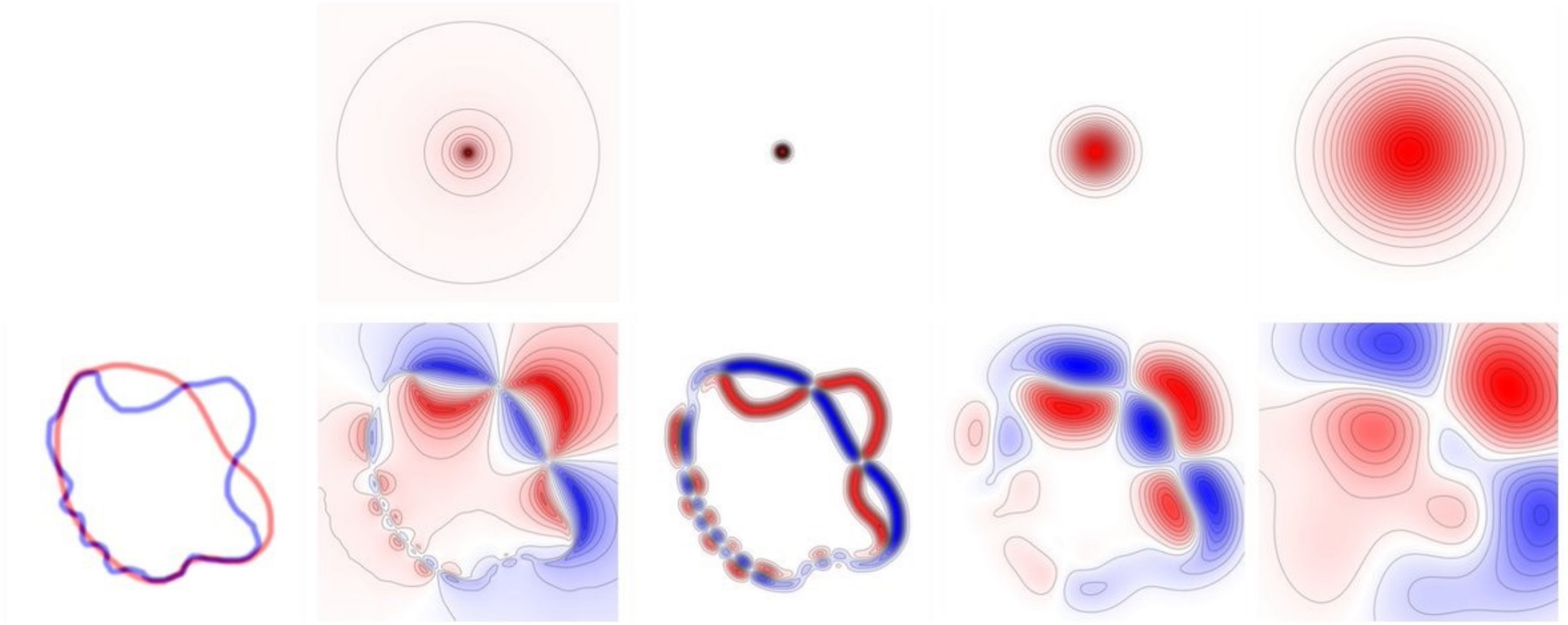
$$\mathcal{D}_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta + \varphi'_\infty \alpha^\perp(\mathcal{X})$$

$$\varphi'_\infty = \lim_{x \uparrow +\infty} \varphi(x)/x \in \mathbb{R} \cup \{\infty\}$$





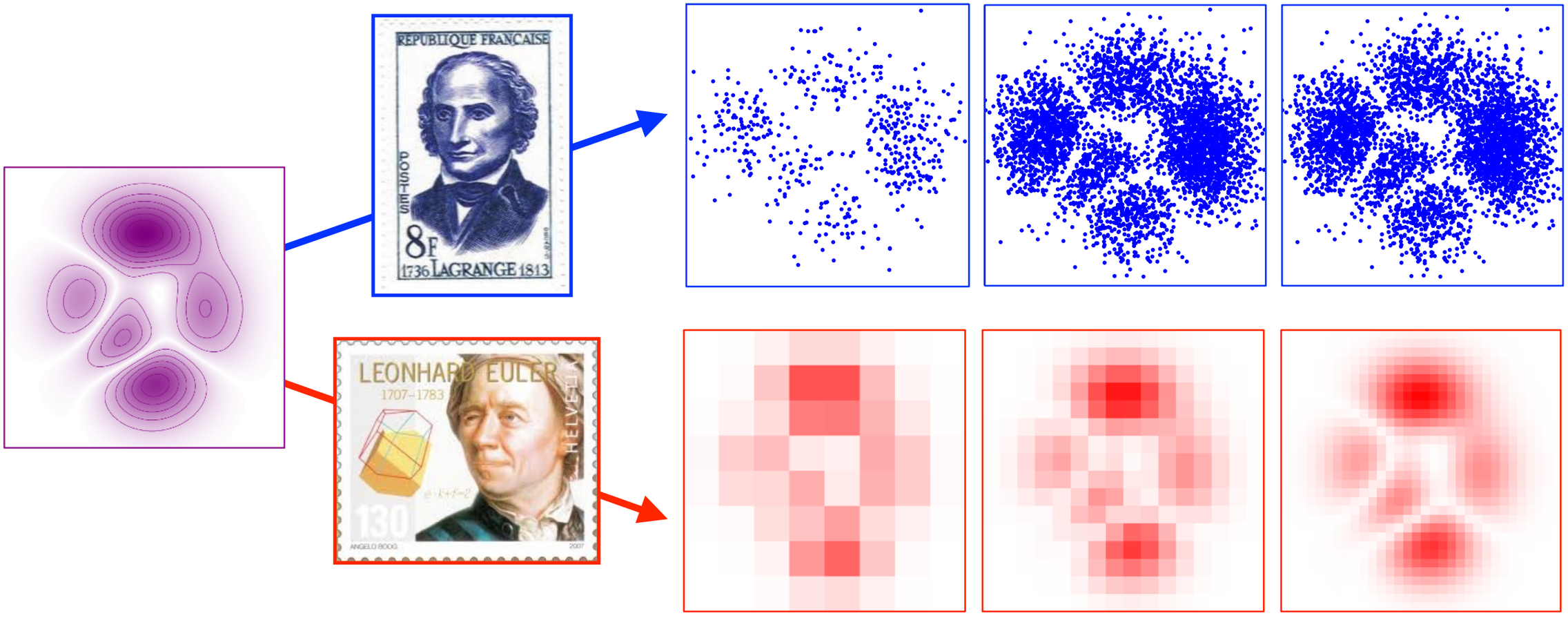
Csiszár divergences, a unifying way to define losses between arbitrary positive measures (discrete & densities). https://en.wikipedia.org/wiki/F-divergence …



(α, β) ED $(G, .005)$ $(G, .02)$ $(G, .05)$

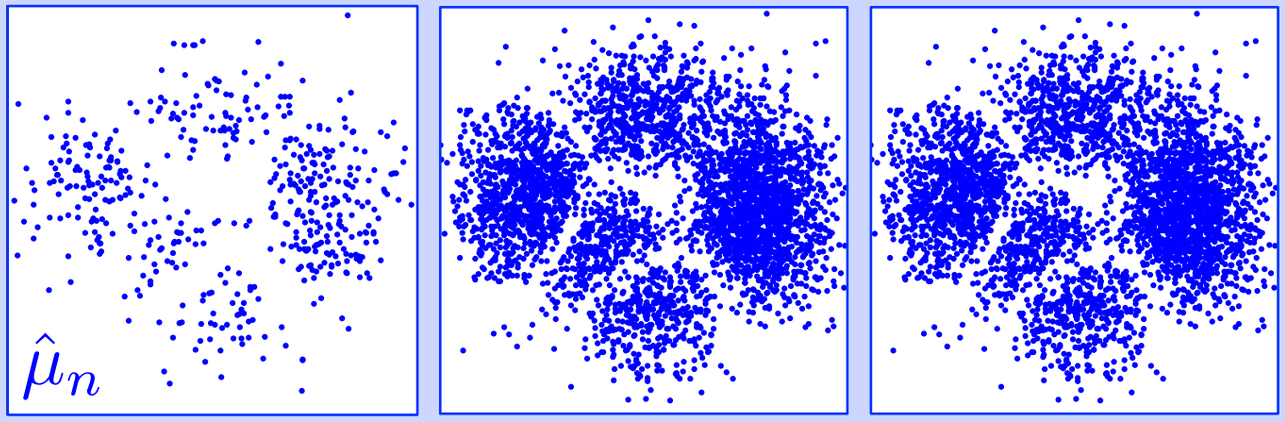
Figure 8.4: Top row: display of ψ such that $\|\alpha - \beta\|_k = \|\psi \star (\alpha - \beta)\|_{L^2(\mathbb{R}^2)}$, formally defined over Fourier as $\hat{\psi}(\omega) = \sqrt{\hat{\varphi}(\omega)}$ where $k^*(x, x') = \varphi(x - x')$. Bottom row: display of $\psi \star (\alpha - \beta)$. (G,$\sigma$) stands for Gaussian kernel of variance σ^2 and ED for Energy Distance kernel (in which case $\psi(x) = 1/\sqrt{\|x\|}$).

Comparing distributions using kernels: heavy tail vs. local (Gaussian) kernels.



Eulerian vs Lagrangian discretization of a probability distribution. Two philosophies of life.

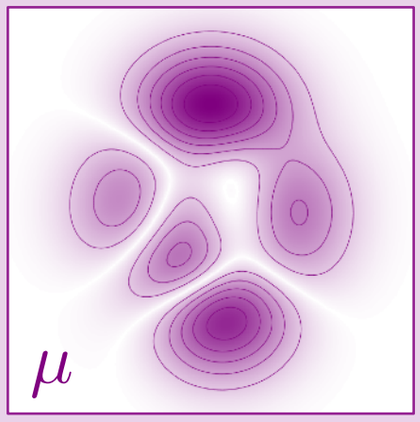
$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$



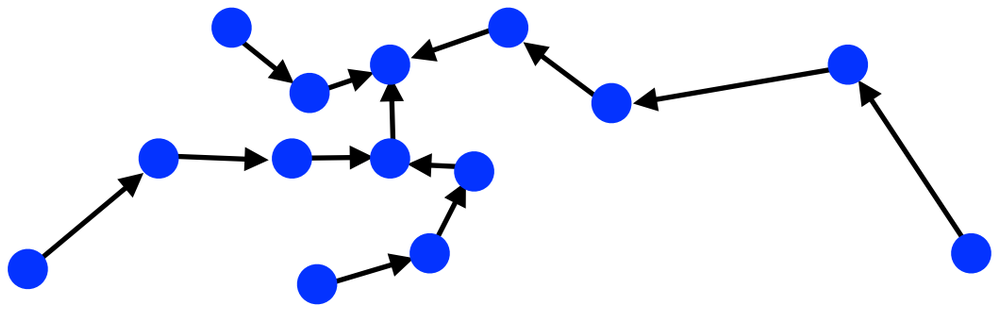
$$\hat{H}(\hat{\mu}_n) \stackrel{\text{def.}}{=} \sum_i \log(\min_{j\neq i} \|x_i - x_j\|)$$

$n \to +\infty$

$n \to +\infty$



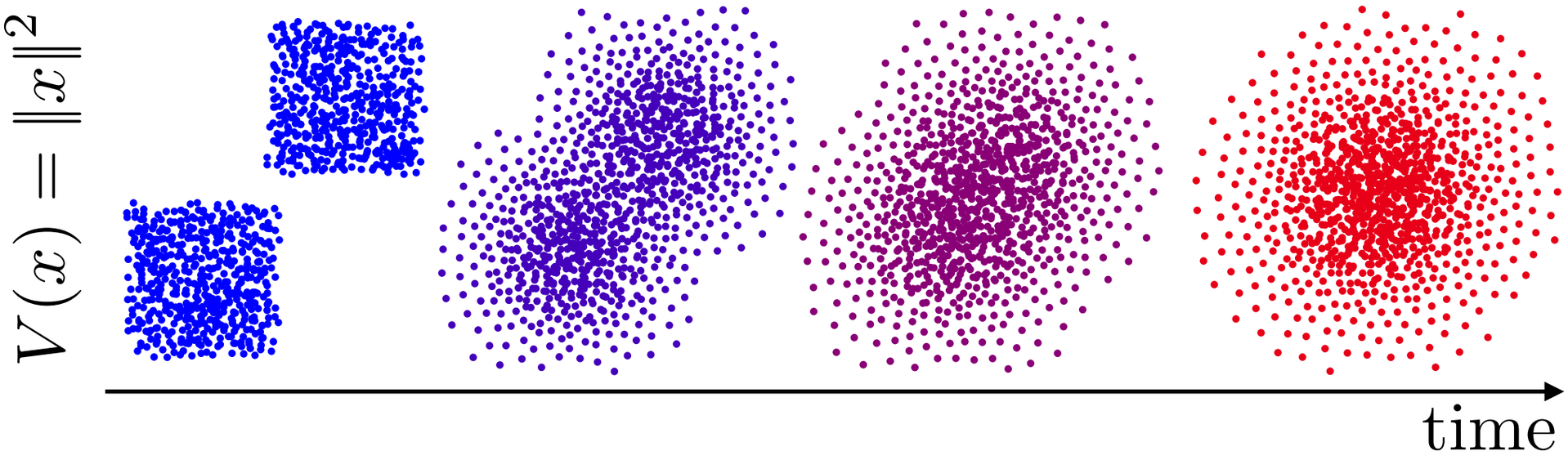
$$H(\mu) \stackrel{\text{def.}}{=} -\int \log(\frac{\mathrm{d}\mu}{\mathrm{d}x}(x))\mathrm{d}\mu(x)$$



Summing the log of the nearest neighbors distances: a simple Lagrangian estimator for the entropy.

$$\min_{\rho} E(\rho) \overset{\text{def.}}{=} \int V(x)\rho(x)\mathrm{d}x + \int \rho(x)\log(\rho(x))\mathrm{d}x$$

Wasserstein flow of E: $\dfrac{\mathrm{d}\rho_t}{\mathrm{d}t} = \Delta\rho_t + \nabla(V\rho_t)$

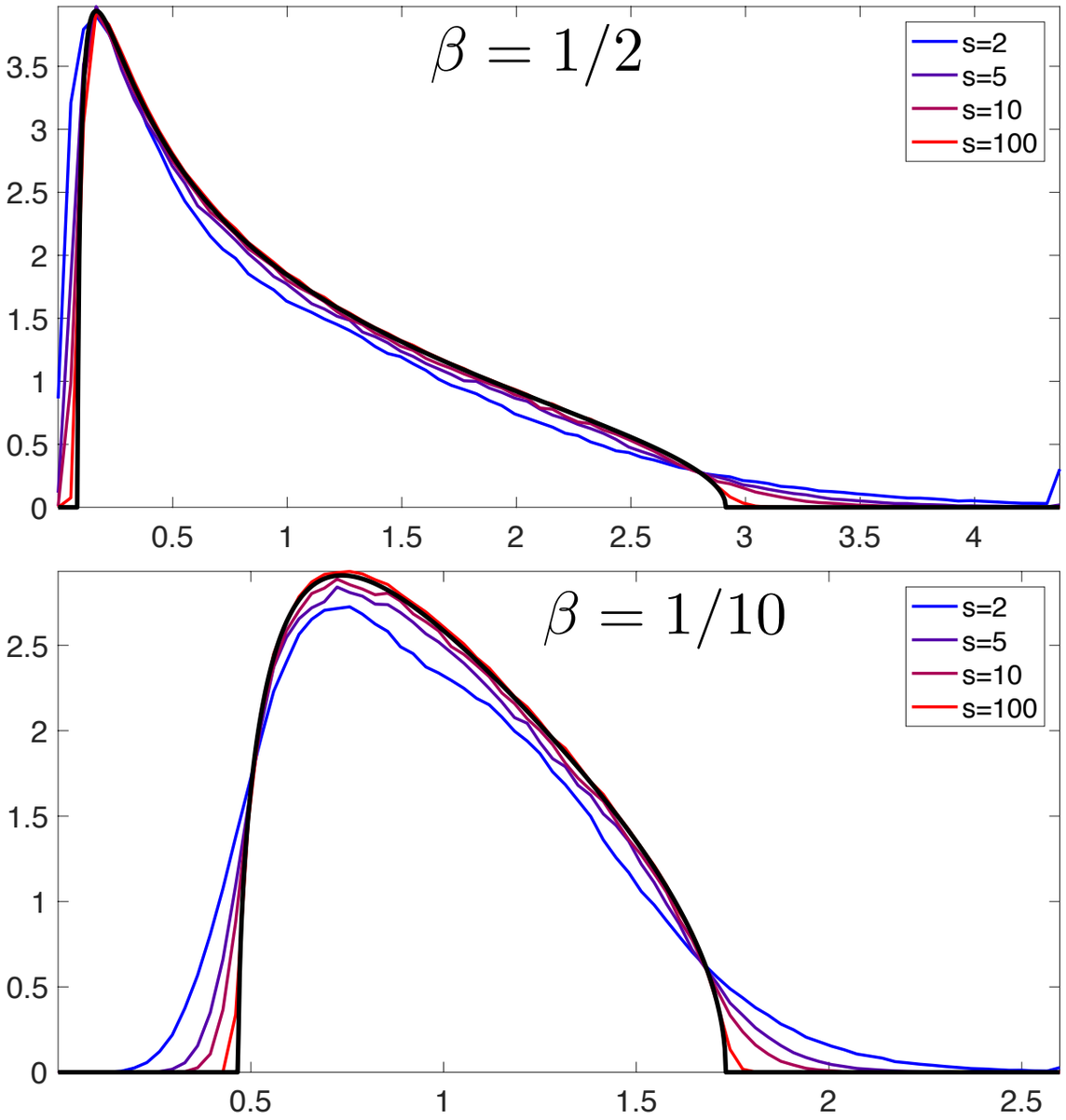


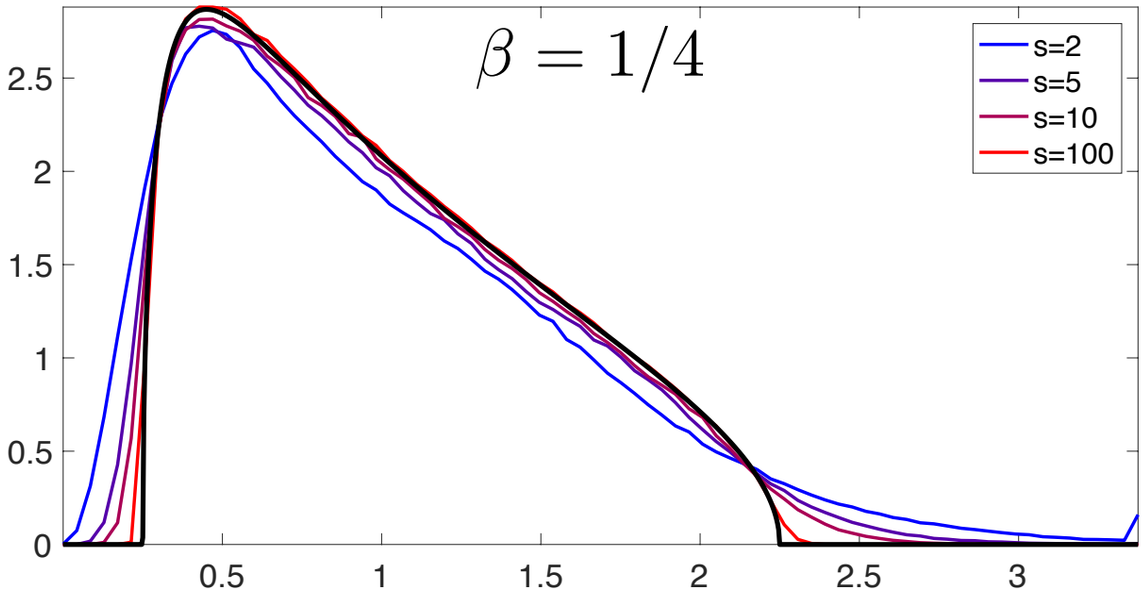
Fokker-Planck equations are Optimal Transport flows of entropies. Nice review by Filippo Santambrogio. https://arxiv.org/abs/1609.03890

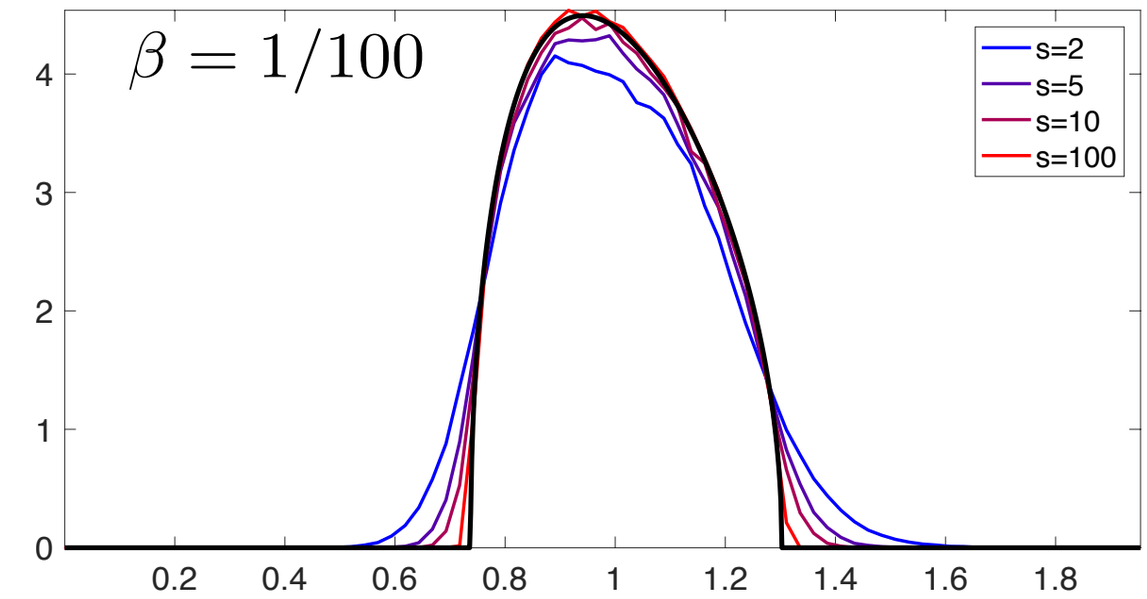
$$B \sim \frac{1}{\sqrt{P}} \texttt{randn}(P, s) \qquad \mathbb{P}(\mathrm{eig}(B^\top B) \in [u, v]) \xrightarrow[s \to +\infty]{\beta \overset{\text{def.}}{=} \frac{s}{P}} \int_u^v f_\beta(\lambda) \mathrm{d}\lambda$$

$$f_\beta(\lambda) \overset{\text{def.}}{=} \frac{1}{2\pi\beta\lambda} \sqrt{(\lambda - \lambda_-)(\lambda_+ - \lambda)} 1_{[\lambda_-, \lambda_+]}(\lambda) \qquad \text{[Marcenko-Pastur]}$$

$$\lambda_\pm \overset{\text{def.}}{=} (1 \pm \sqrt{\beta})^2$$

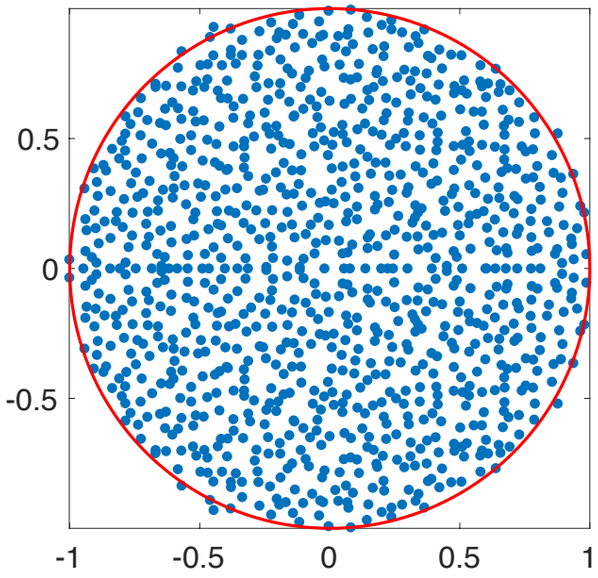






The Marchenko-Pastur law is the distribution limit of eigenvalues of covariance matrices when samples and size grow together. http://djalil.chafai.net/blog/2011/01/29/the-marchenko-pastur-

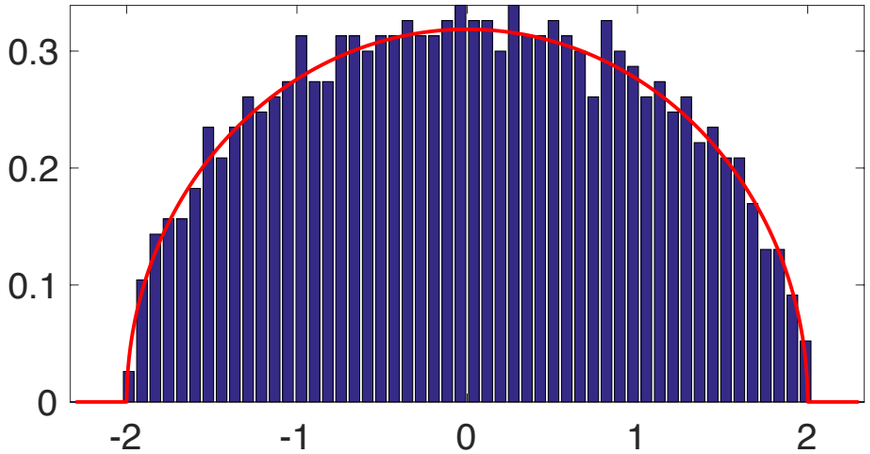
$\mathrm{eig}(\mathtt{randn}(N,N))$



Circle law

$A = \mathtt{randn}(N)$

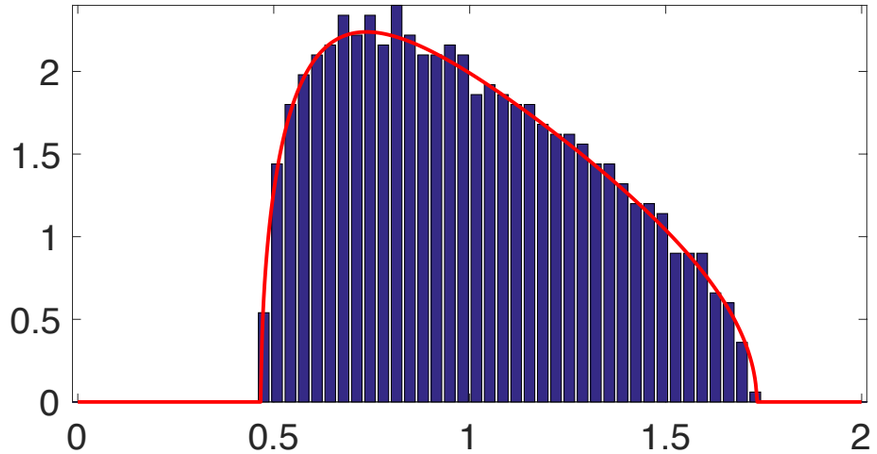
$\mathrm{eig}(A + A^\top)$



Semi-circle law

$B = \mathtt{randn}(N, N/10)$

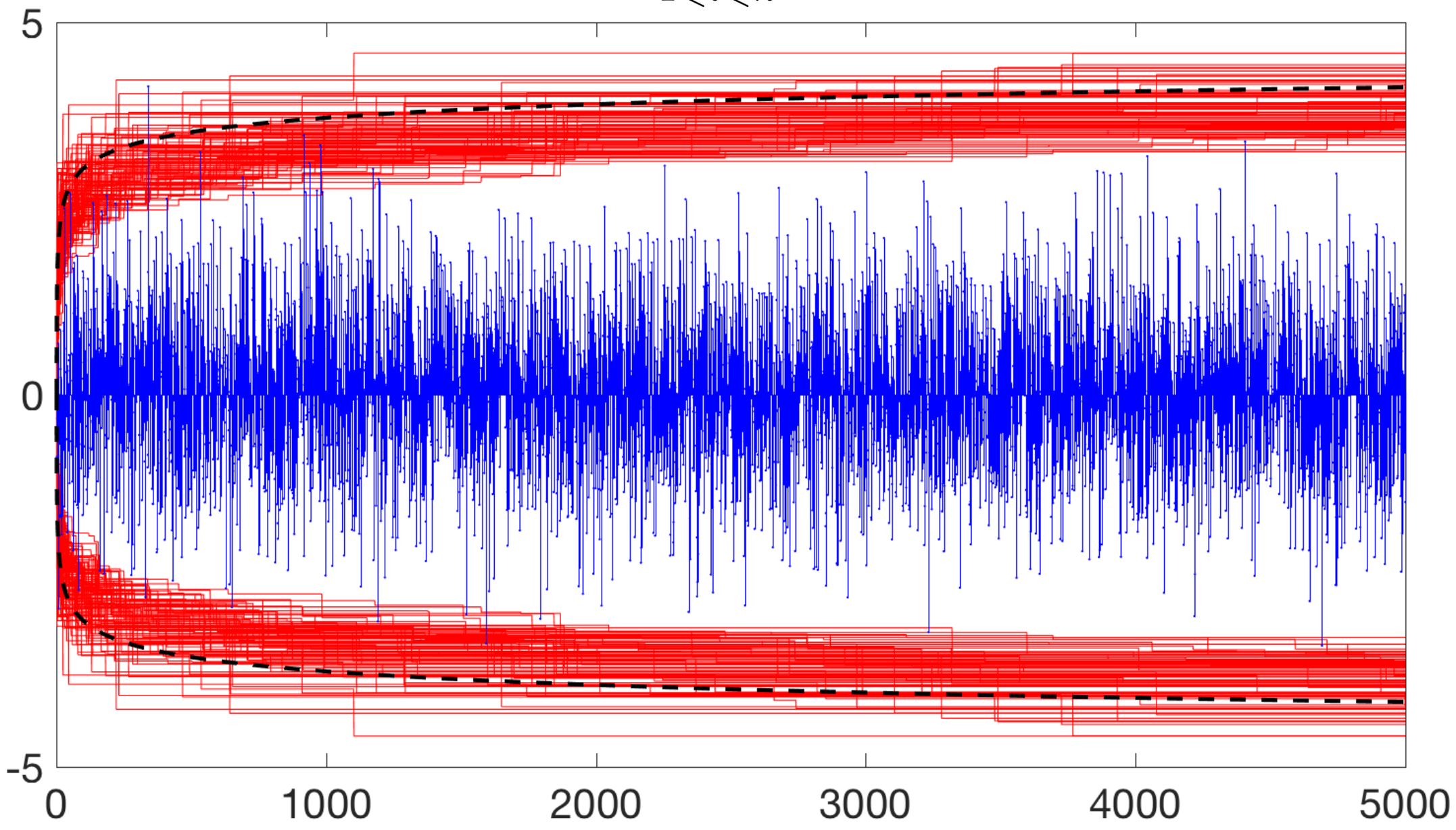
$\mathrm{eig}(B^\top B)$



Marcenko Pastur law

The circle, semi-circle and Marčenko–Pastur laws are three simple examples of determinantal processes. https://terrytao.wordpress.com/2009/08/23/determinantal-processes/ …

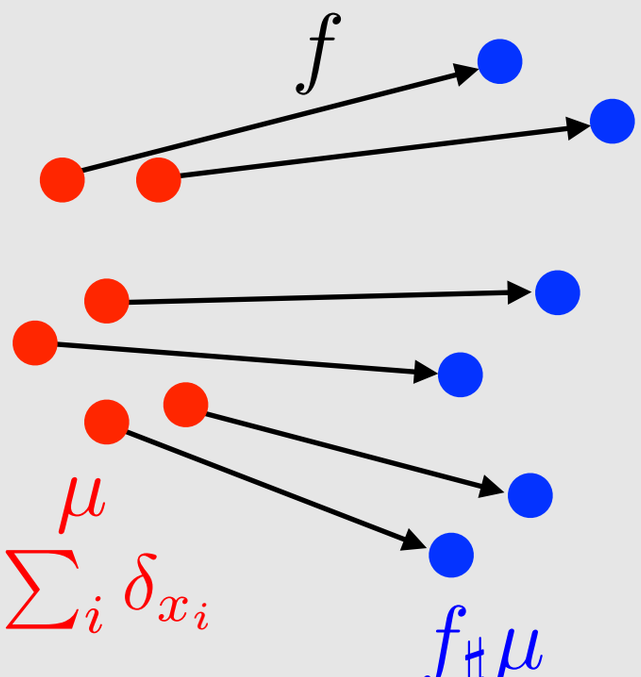
$$X_i \sim \mathcal{N}(0,1) \qquad \max_{1 \leqslant i \leqslant n} |X_i| \sim \sqrt{2\log(n)}$$



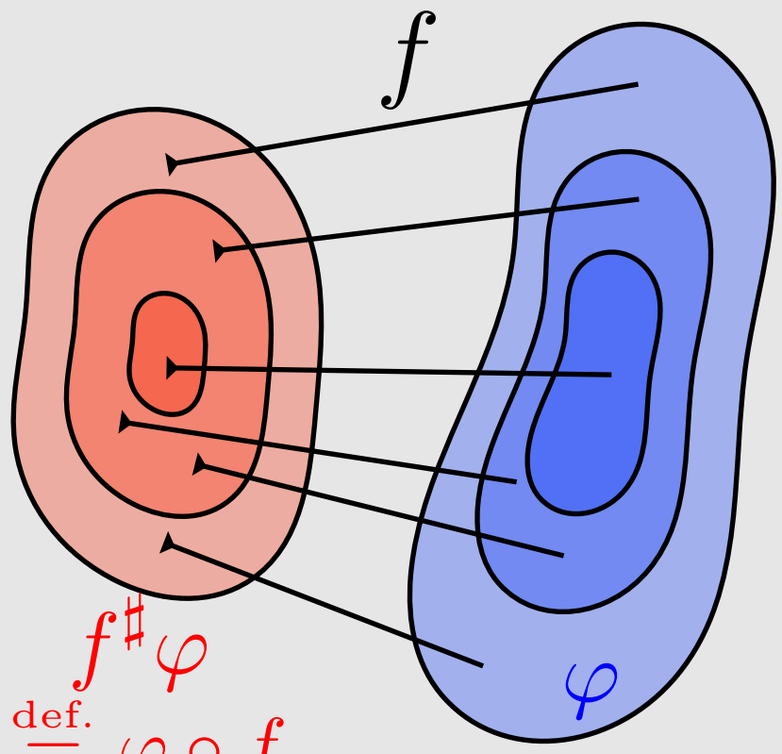
The maximum of n i.i.d. Gaussians is roughly sqrt(2log(n)). This is where log factors come from in Compressed Sensing.

$f : \mathcal{X} \to \mathcal{Y}$

Measures: push-forward



Functions: pull-back



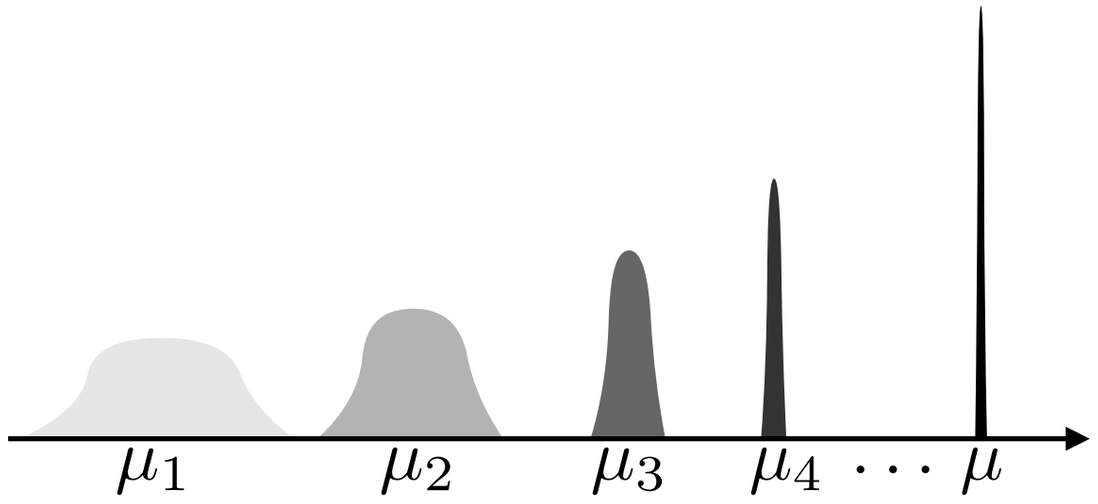
$$f_\sharp : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y}) \qquad f^\sharp : \mathcal{C}(\mathcal{Y}) \to \mathcal{C}(\mathcal{X})$$

Remark: $f^\sharp$ and $f_\sharp$ are adjoints

$$\int_{\mathcal{Y}} \varphi \mathrm{d}(f_\sharp \mu) = \int_{\mathcal{X}} (f^\sharp \varphi) \mathrm{d}\mu$$

Functions should be manipulated by pullback, measures by pushforward. These are adjoints operations.

Random vectors		Radon measures
$\mathbb{P}(X \in A)$	$=$	$\int_A \mathrm{d}\mu(x)$
Weak* convergence:		Convergence in law:
$\forall$ set A $\quad \mathbb{P}(X_n \in A) \overset{n\to+\infty}{\longrightarrow} \mathbb{P}(X \in A)$		$\forall$ continuous function f $\quad \int f \mathrm{d}\mu_n \overset{n\to+\infty}{\longrightarrow} \int f \mathrm{d}\mu$



Convergence in law of random vectors and
weak* convergence of measures are the same. #LetsBeFriends
https://en.wikipedia.org/wiki/Convergence_of_measures
https://en.wikipedia.org/wiki/Convergence_of_random_variables

In mean

$$\lim_{n\to+\infty} \mathbb{E}(|X_n - X|^p) = 0$$

Almost sure

$$\mathbb{P}(\lim_{n\to+\infty} X_n = X) = 1$$

In probability

$$\forall \varepsilon > 0,\ \mathbb{P}(|X_n - X| > \varepsilon) \overset{n\to+\infty}{\longrightarrow} 0$$

In law

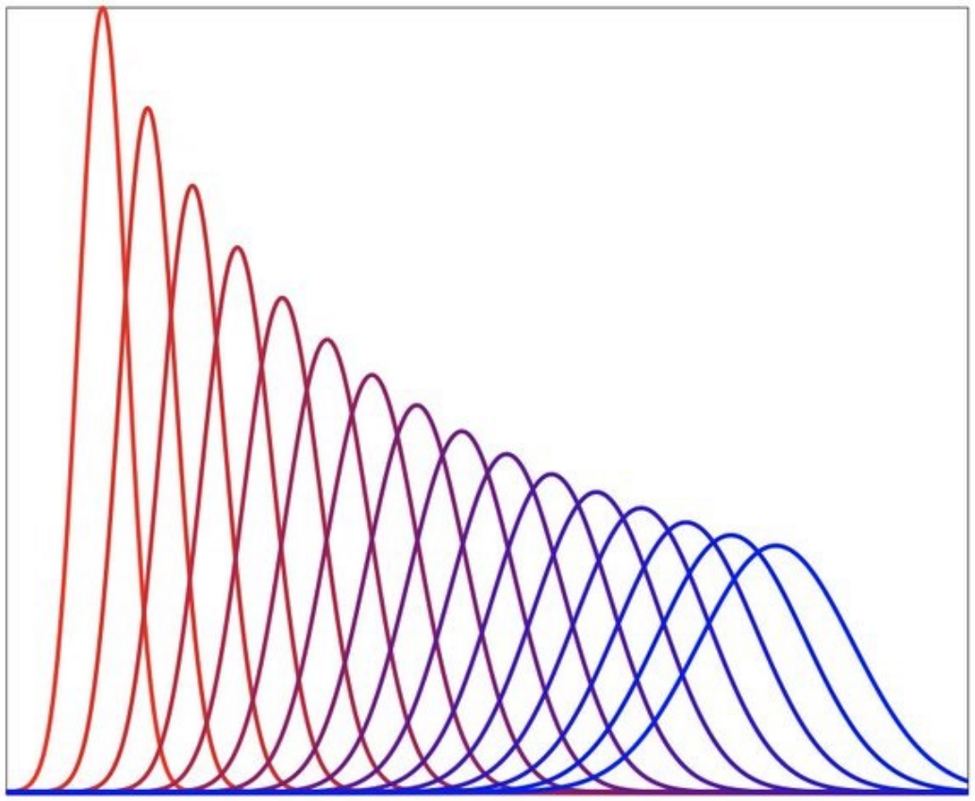
$$\mathbb{P}(X_n \in A) \overset{n\to+\infty}{\longrightarrow} \mathbb{P}(X \in A)$$

(the X_n can be defined on different spaces)

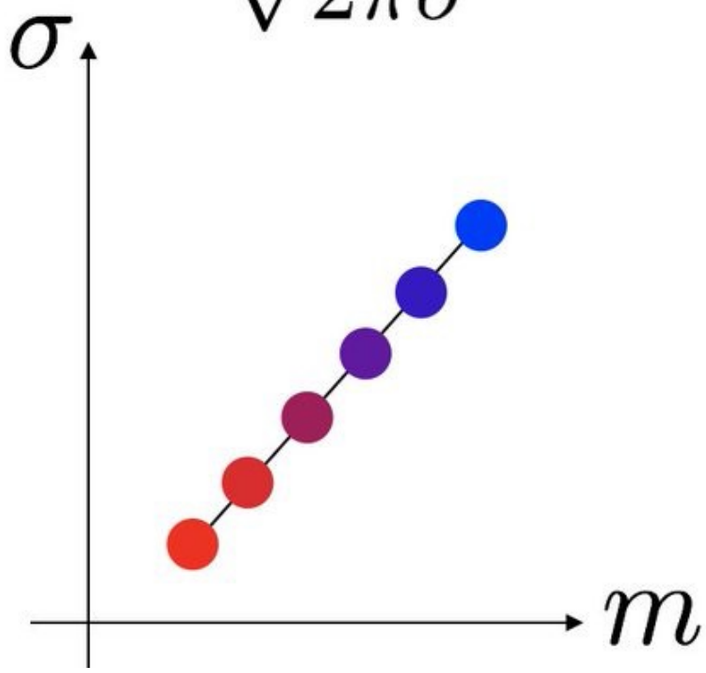
Convergence of random vectors comes with lots of flavors!

https://en.wikipedia.org/wiki/Convergence_of_random_variables

Optimal Transport



$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$



The Optimal Transport geometry of 1-D Gaussians is flat in the (mean,std) plane. #DoesNotWorkIn2D

Dual norms: (aka Integral Probability Metrics)

$$\|\alpha - \beta\|_B \overset{\text{def.}}{=} \max\left\{ \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x)) \;;\; f \in B \right\}$$

Wasserstein 1: $B = \{f \;;\; \|\nabla f\|_\infty \leqslant 1\}$.

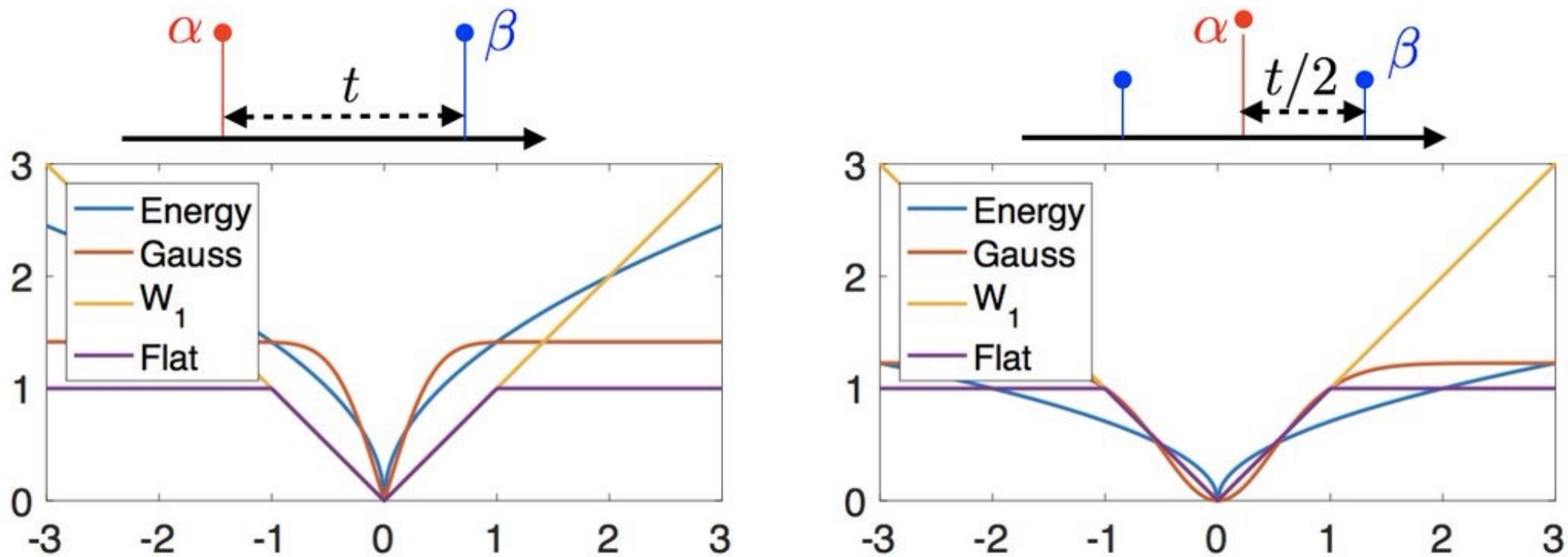
Flat norm: $B = \{f \;;\; \|f\|_\infty \leqslant 1, \|\nabla f\|_\infty \leqslant 1\}$.

RKHS: $B = \left\{f \;;\; \|f\|_k^2 \leqslant 1\right\}$.

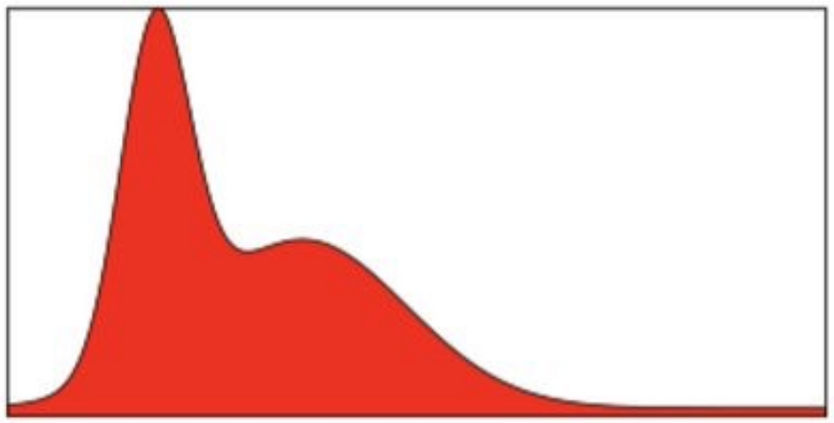
$$\|\alpha - \beta\|_B^2 = \int k(x,x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int k(x,x')\mathrm{d}\beta(y)\mathrm{d}\beta(y') - 2\int k(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$$

Energy distance: $k(x,y) = -\|x-y\|$

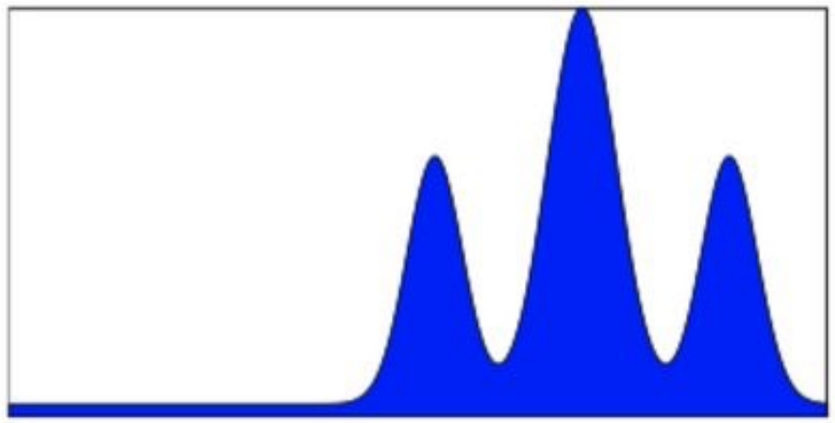
Gaussian: $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$



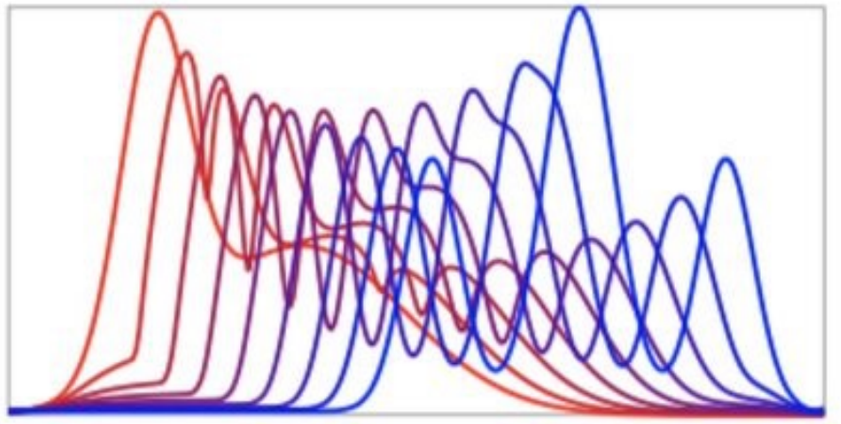
Dual weak norms (aka Integral Probability Metrics): a unifying way to define norms over measures compatible with weak convergence.



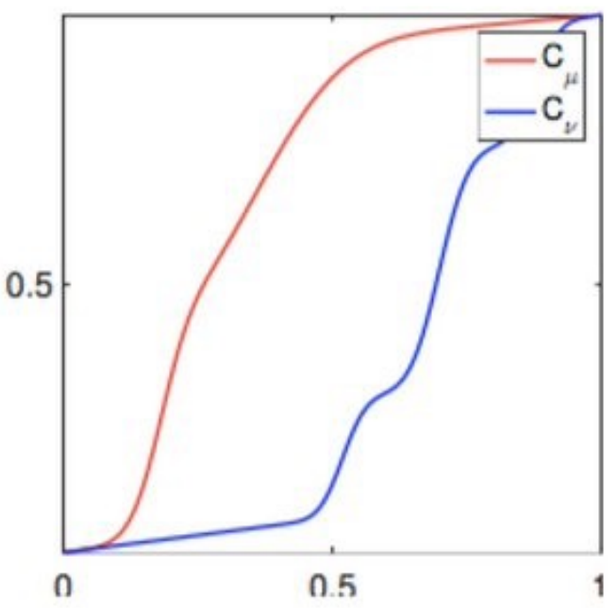
α



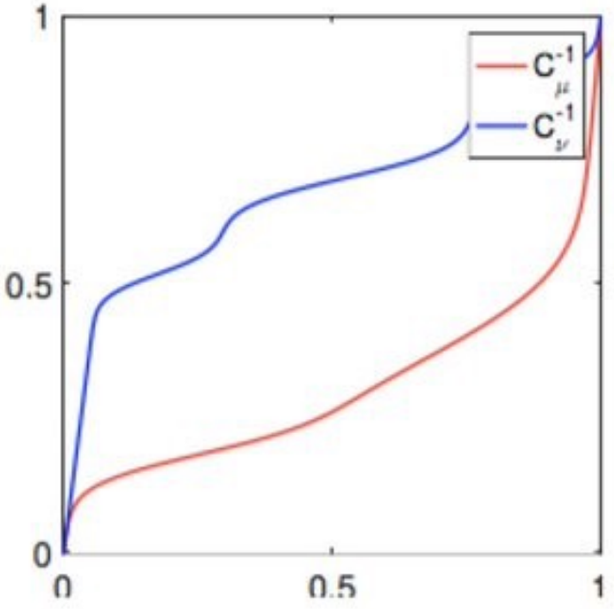
β



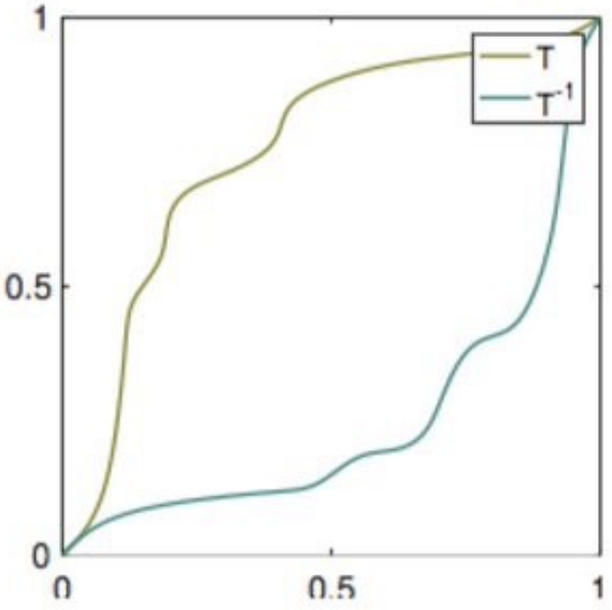
$((t\mathrm{Id} + (1-t)\mathrm{Id})_\sharp \alpha$



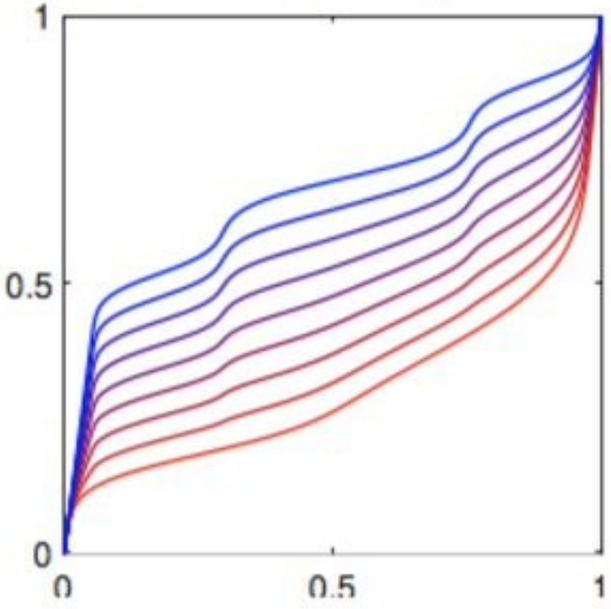
$(\mathcal{C}_\alpha, \mathcal{C}_\beta)$



$(\mathcal{C}_\alpha^{-1}, \mathcal{C}_\beta^{-1})$



(T, T^{-1})



$(1-t)\mathcal{C}_\alpha^{-1} + t\mathcal{C}_\beta^{-1}$

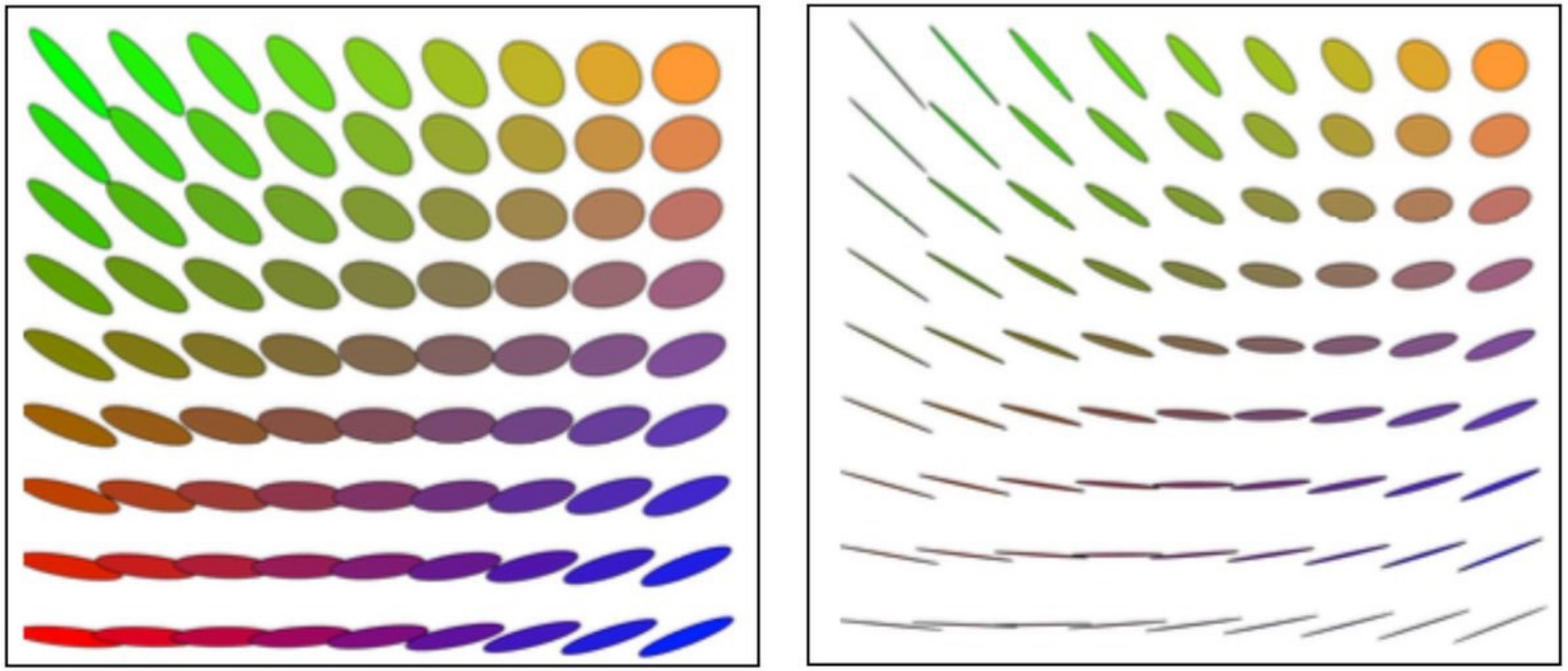
Displacement interpolation aka Optimal Transport in 1D is equivalent to interpolating the *inverse* cumulative functions.

Remark 2.11 (Distance between Gaussians). If $\alpha = \mathcal{N}(m_\alpha, C_\alpha)$ and $\beta = \mathcal{N}(m_\beta, C_\beta)$, then one can show that

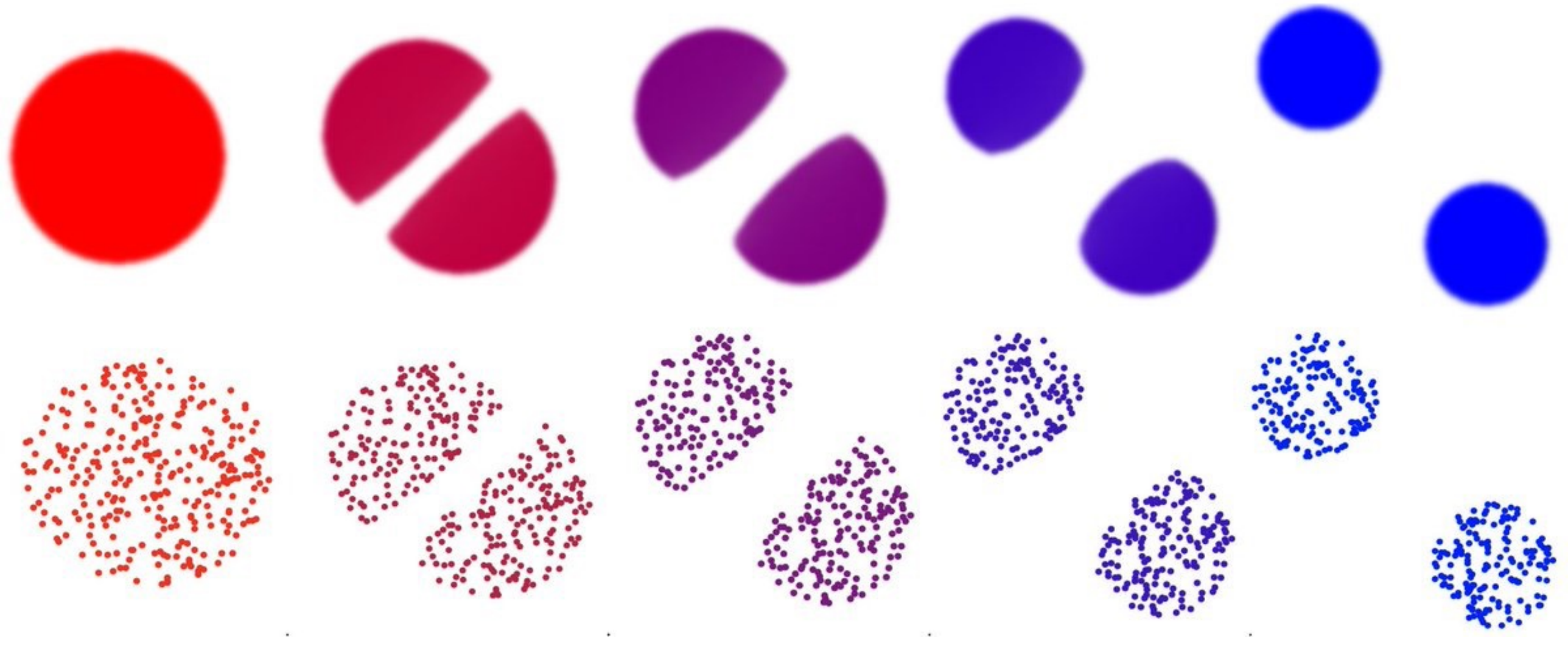
$$\mathcal{W}_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}(C_\alpha, C_\beta)^2 \tag{2.19}$$

where $\mathcal{B}$ is the so-called Bures metric

$$\mathcal{B}(C_\alpha, C_\beta)^2 \overset{\text{def.}}{=} \operatorname{tr}\left(C_\alpha + C_\alpha - 2(C_\alpha^{1/2} C_\beta C_\alpha^{1/2})^{1/2}\right) \tag{2.20}$$

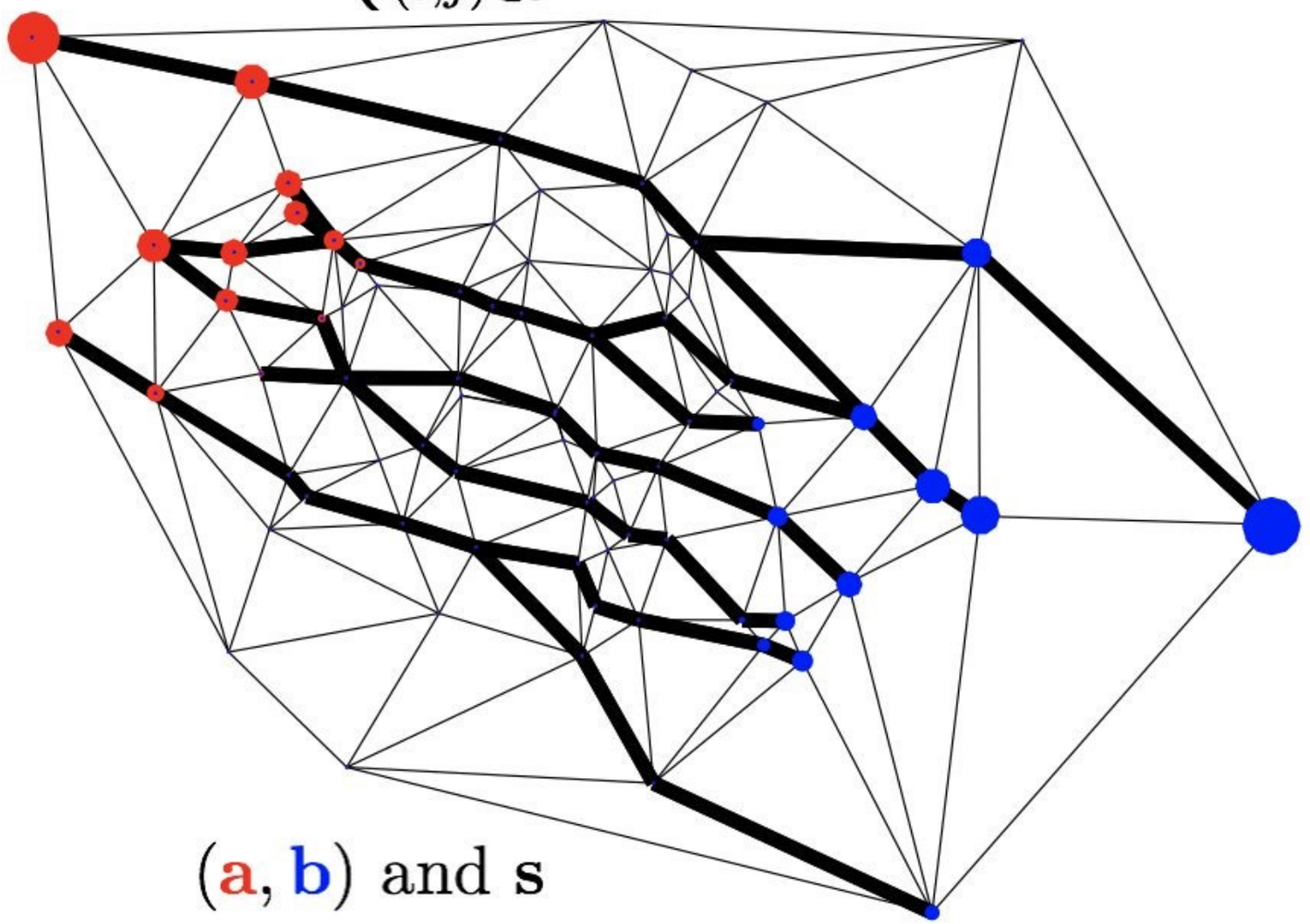


Bures distance on PSD matrices, aka Optimal Transport between Gaussians, handles nicely singular matrices https://en.wikipedia.org/wiki/Bures_metric …

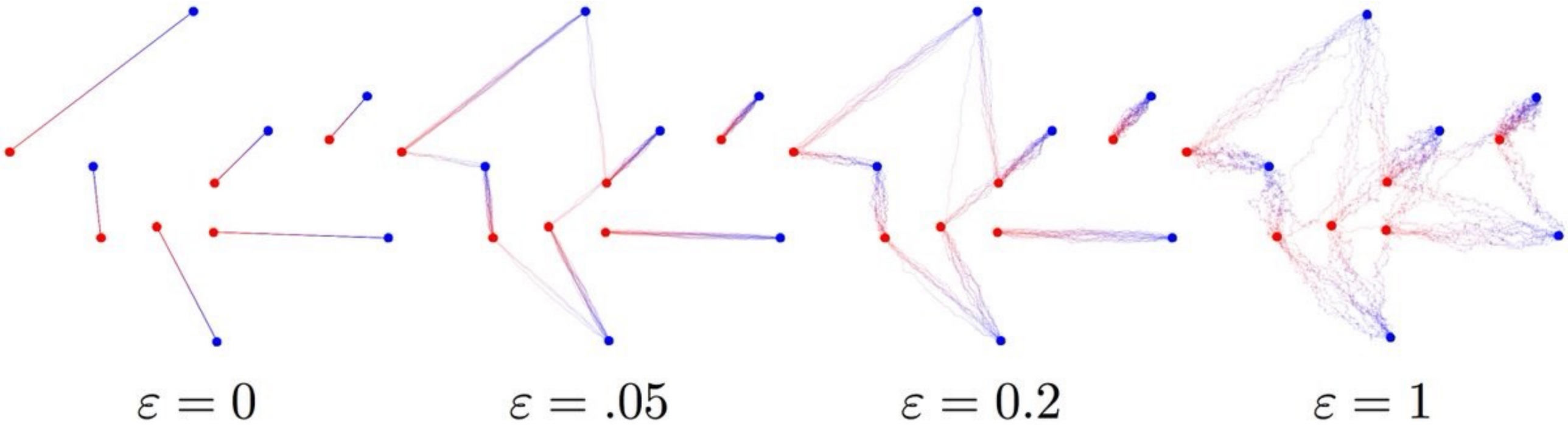


Displacement interpolation aka Optimal Transport: discrete vs continuous.

$$\mathrm{W}_1(\mathbf{a},\mathbf{b}) = \min_{\mathbf{s}\in\mathbb{R}_+^{\mathcal{E}}} \left\{ \sum_{(i,j)\in\mathcal{E}} \mathbf{w}_{i,j}\mathbf{s}_{i,j} \;:\; \mathrm{div}(\mathbf{s}) = \mathbf{a}-\mathbf{b} \right\}$$



Wasserstein-1 distance (norm!) between measures on graphs: equivalent to a min-cost flow. Bold black indicates edges where mass is flowing.



$\varepsilon = 0$ $\varepsilon = .05$ $\varepsilon = 0.2$ $\varepsilon = 1$

From deterministic to stochastic matching: Schrödinger problem.
See Leonard's survey for the maths. https://arxiv.org/abs/1308.0215

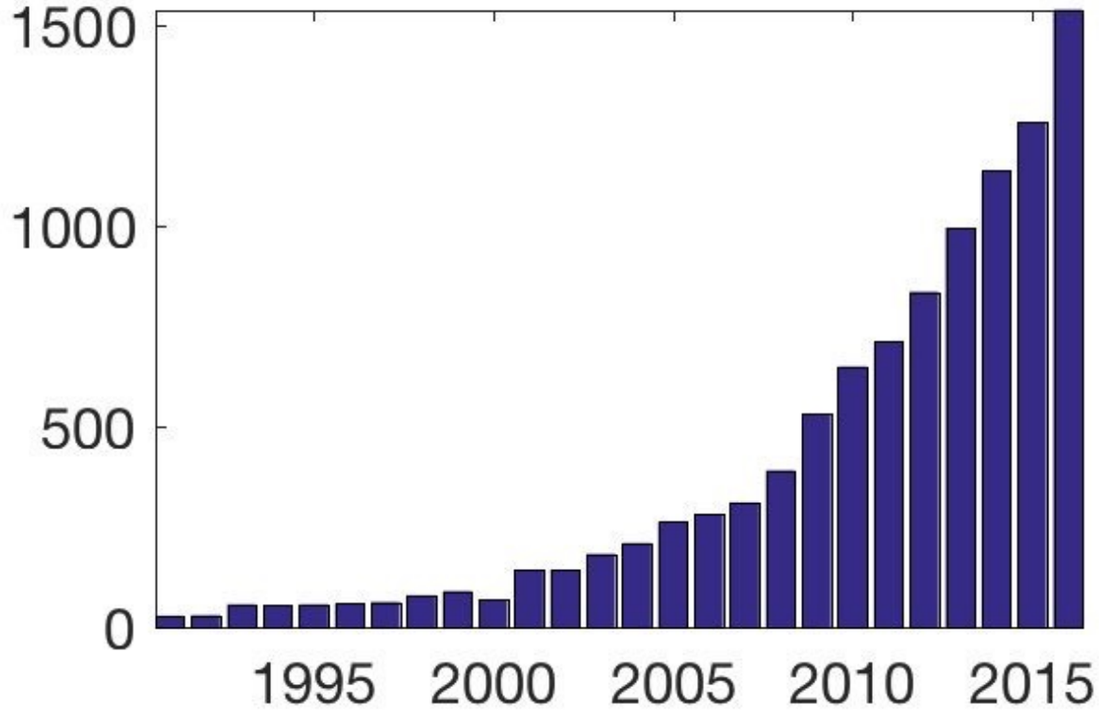
If (μ, ν) have densities bounded from below/above by $0 < a < b < +\infty$,

$$b^{-1/2}\|\mu - \nu\|_{H^{-1}} \leqslant W_2(\mu, \nu) \leqslant a^{-1/2}\|\mu - \nu\|_{H^{-1}}$$

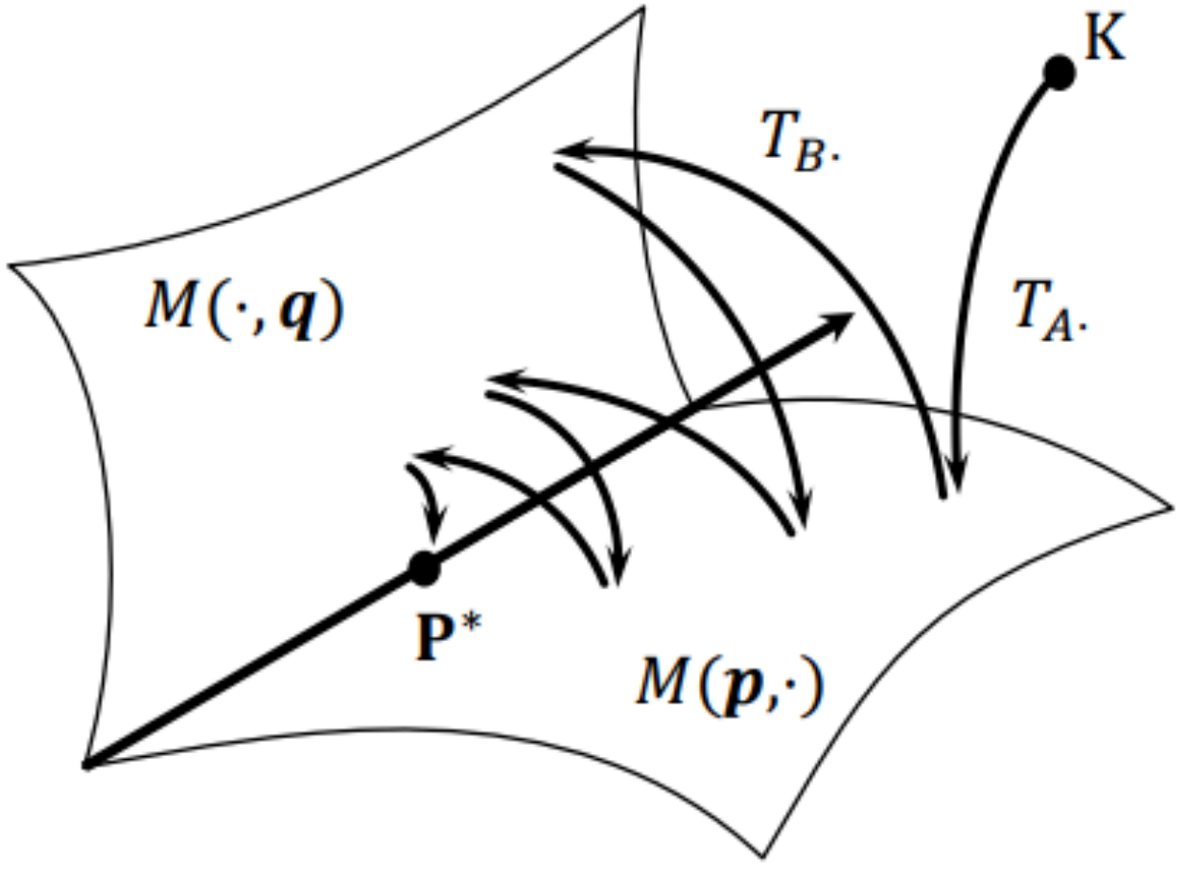
Wasserstein-2 and Energy Distance (Sobolev H^-1) are equivalent for bounded densities. Sharp constant by R. Peyre https://arxiv.org/abs/1104.4631v2

$$W_c(\mu, \nu) = \sup_{f,g} \left\{ \int_X f \mathrm{d}\mu + \int_Y g \mathrm{d}\nu \; ; \; \forall (x, y), f(x) + g(y) \leqslant c(x, y) \right\}$$
$$W_c^*(f, g) = \inf_{x,y} \; c(x, y) - f(x) + g(y)$$

The Legendre transform of the Wasserstein distance has a nice expression! #ProbablyUseless



Google Scholar citations for "Optimal Transport". Soon trendier than Deep Learning!



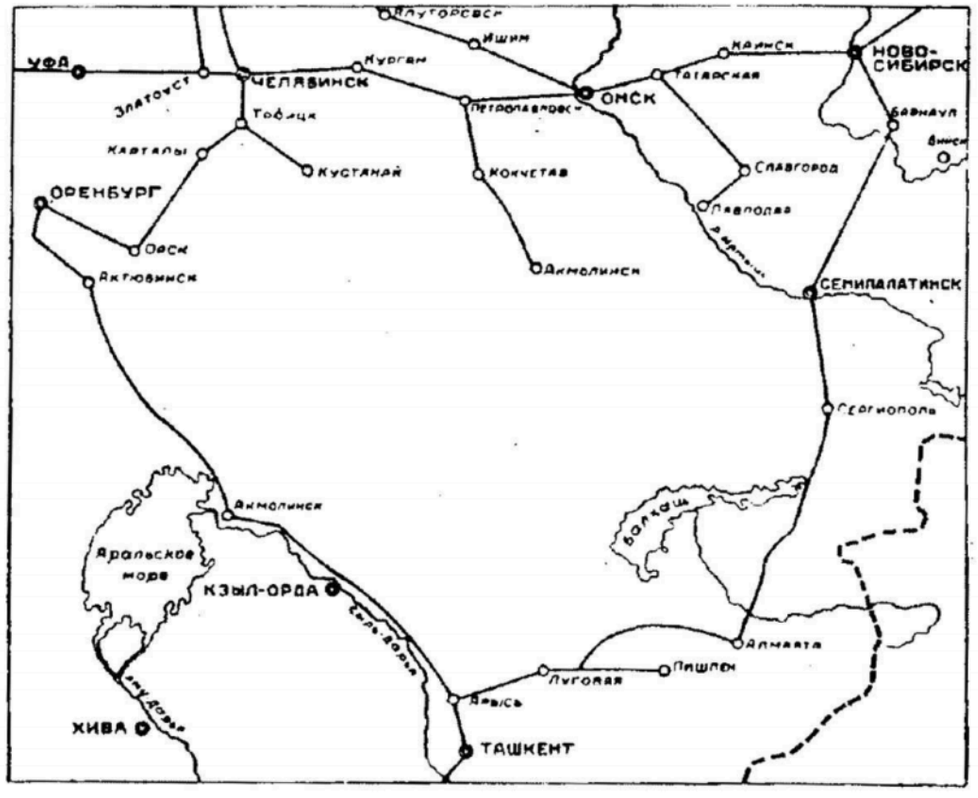
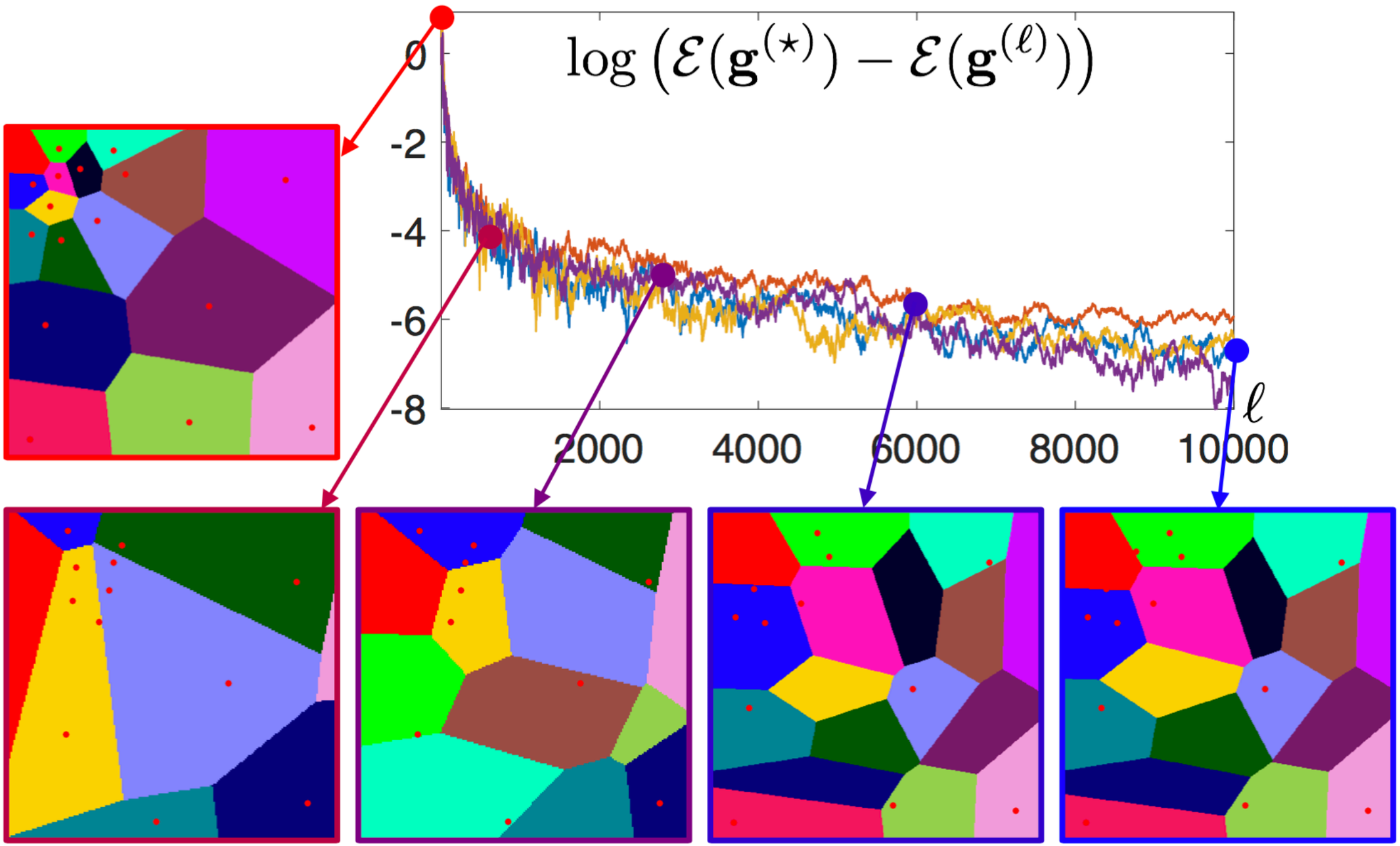
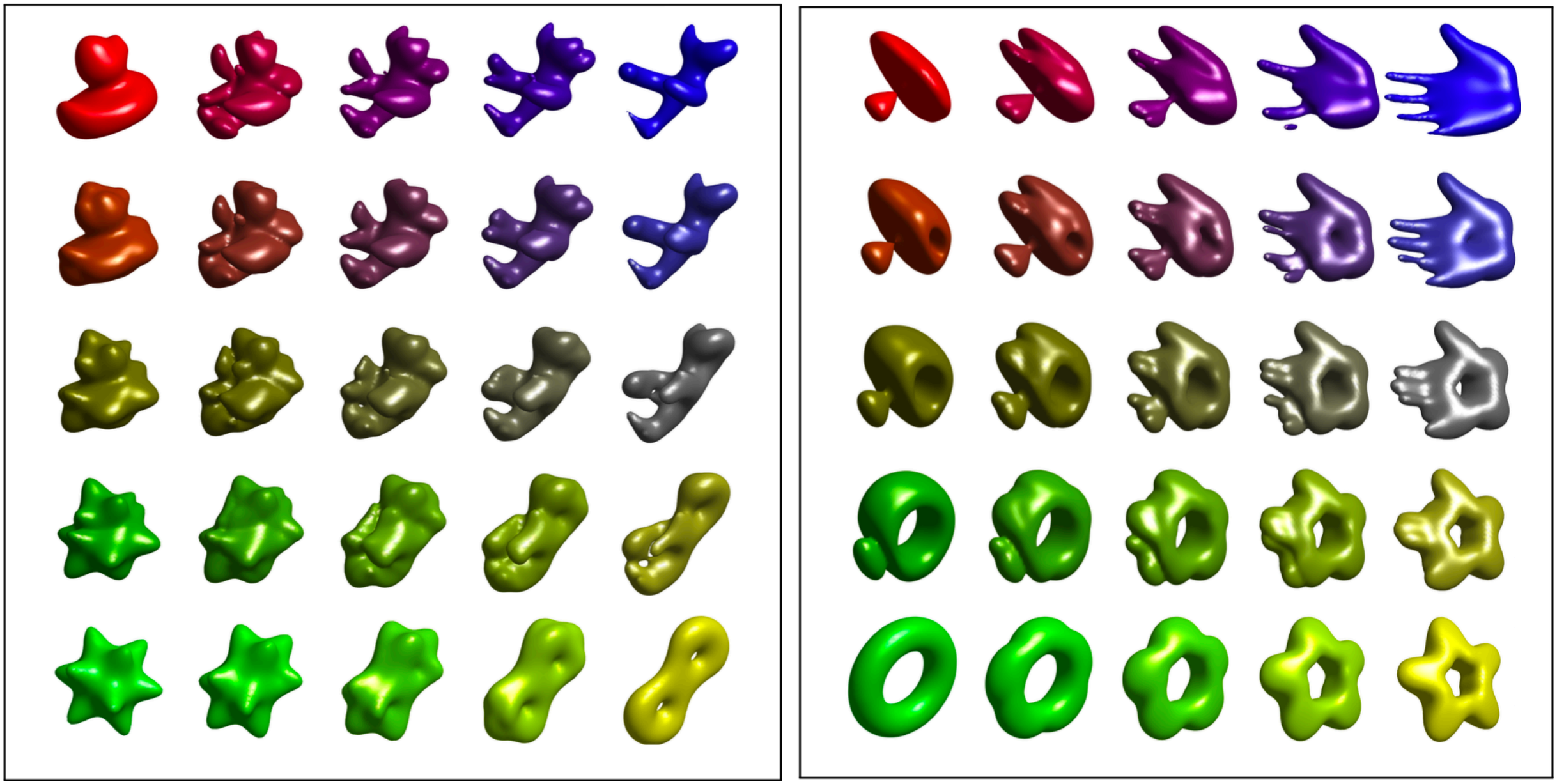


Figure 1: Figure from Tolstoĭ [1930] to illustrate a negative cycle

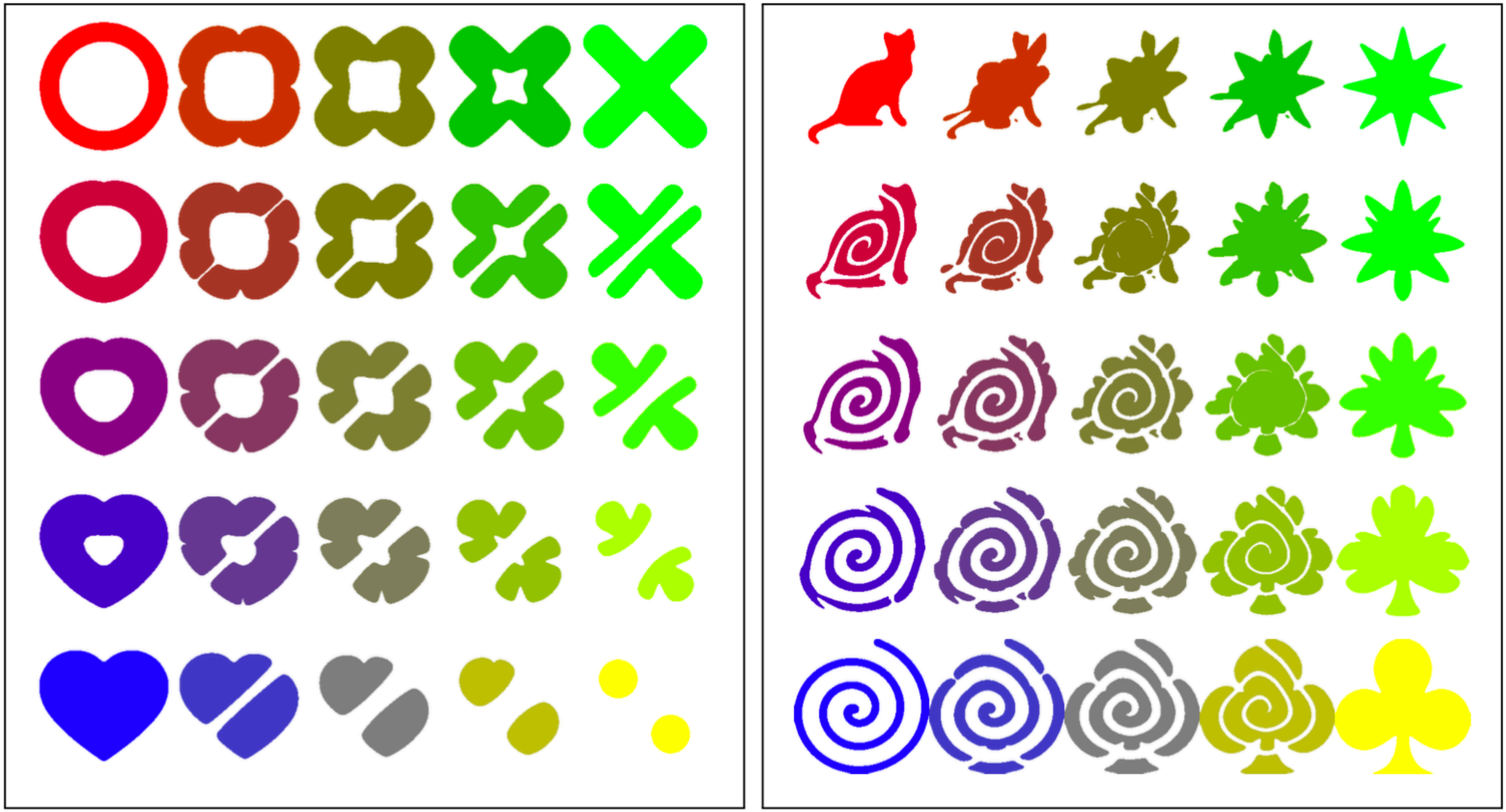
Optimal Transport was formulated in 1930 by A.N. Tolstoi, 12 years before Kantorovich. He even solved a "large scale" 10×68 instance!



Stochastic gradient descent for the semi-discrete Optimal Transport, illustration of convergence and corresponding Laguerre cells. https://arxiv.org/abs/1605.08527



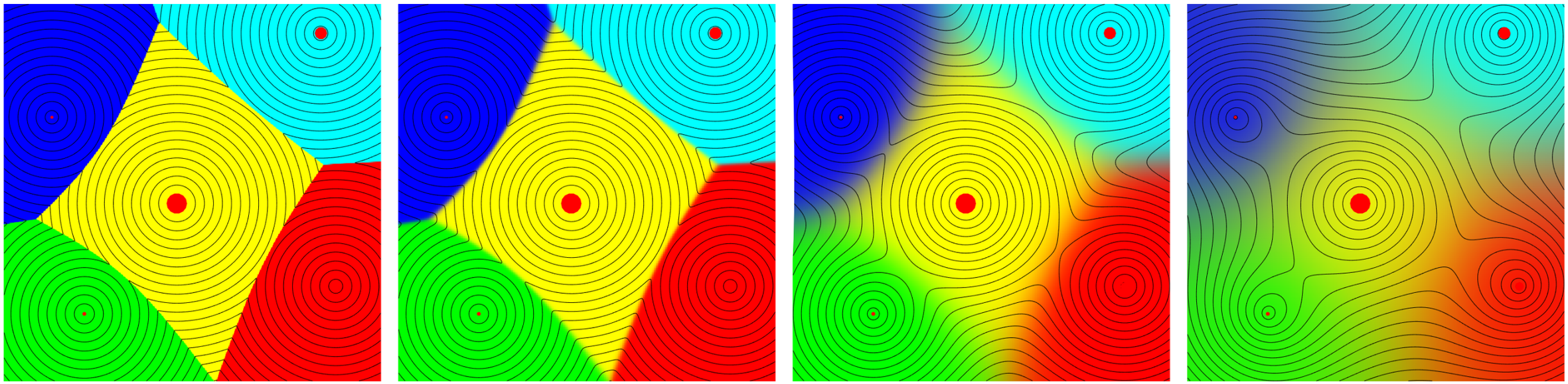
Optimal Transport barycenters. https://hal.archives-ouvertes.fr/hal-01188953 https://hal.archives-ouvertes.fr/hal-01096124



Heartbroken by Wasserstein barycenters!

Laguerre cells

"Sinkhorn" Laguerre cells



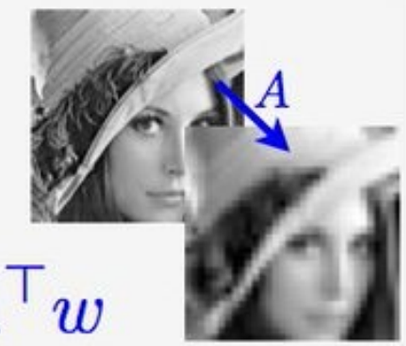
$\varepsilon = 0$ $\varepsilon = 0.01$ $\varepsilon = 0.1$ $\varepsilon = 0.3$

Semi-discrete Optimal Transport is a multiclass SVM. Sinkhorn version is logistic regression. Decomposes space in entropic Laguerre diagrams

Machine Learning & Inverse Problems

Inverse Problems

$$y = Af + w$$



$$A^\top y = (A^\top A) f + A^\top w$$
$$\overset{\text{def.}}{=} u \qquad \overset{\text{def.}}{=} C \qquad \overset{\text{def.}}{=} r$$

Regularized inversion:

$$\min_f \frac{1}{2}\|Af - y\|^2 + \lambda \|f\|^2$$

$$f_\lambda = (C + \lambda \mathrm{Id}_p)^{-1} u$$

Statistical Learning

$$y = Xf + \varepsilon$$

$$\frac{1}{n} X^\top y = \frac{1}{n}(X^\top X) f + \frac{1}{n} X^\top \varepsilon$$
$$\overset{\text{def.}}{=} u_n \qquad \overset{\text{def.}}{=} C_n \qquad \overset{\text{def.}}{=} r_n$$
$$\downarrow n \to +\infty \qquad \downarrow (x_i, y_i)_i \text{ i.i.d.}$$
$$u = \mathbb{E}(yx) \qquad C = \mathbb{E}(xx^\top)$$

Empirical risk minimization:

$$\min_f \frac{1}{2n}\|Xf - y\|^2 + \lambda \|f\|^2$$

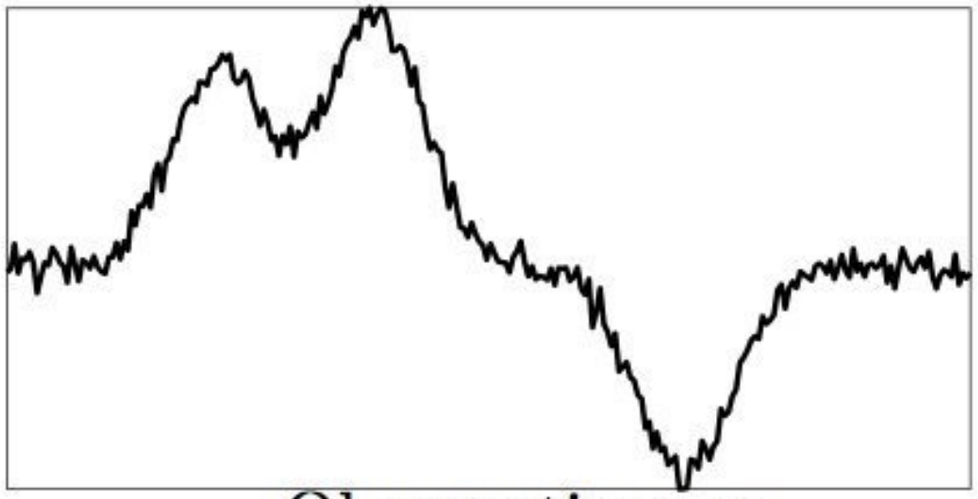
$$f_{\lambda,n} = (C_n + \lambda \mathrm{Id}_p)^{-1} u_n$$

Exact covariance C ⟷ Noisy covariance C_n

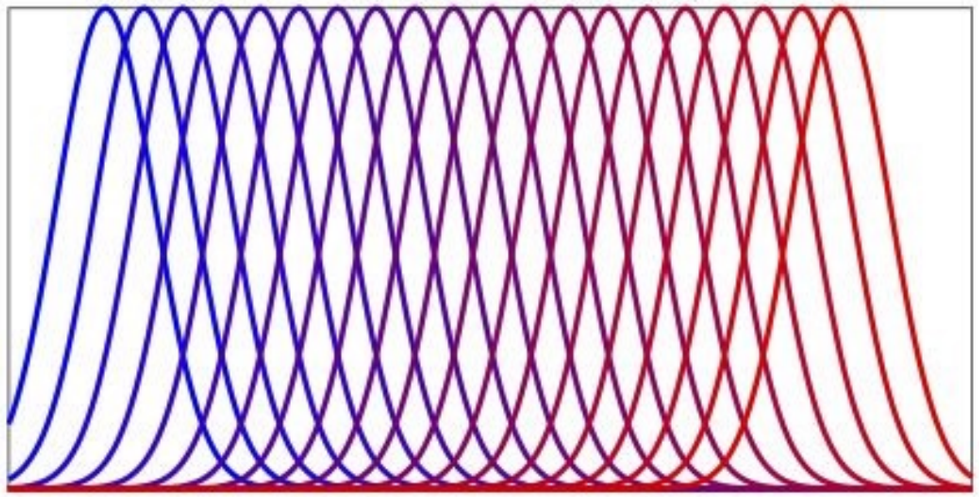
Deterministic bounded noise r ⟷ Random noise r_n

Noise level $\varepsilon \overset{\text{def.}}{=} \|r\|$ ⟷ Noise level $\|r_n\| \sim \varepsilon = n^{-\frac{1}{2}}$

Statistical Learning is Inverse Problems with a noisy covariance matrix / IP is SL
with a deterministic covariance matrix #LetsBeFriends

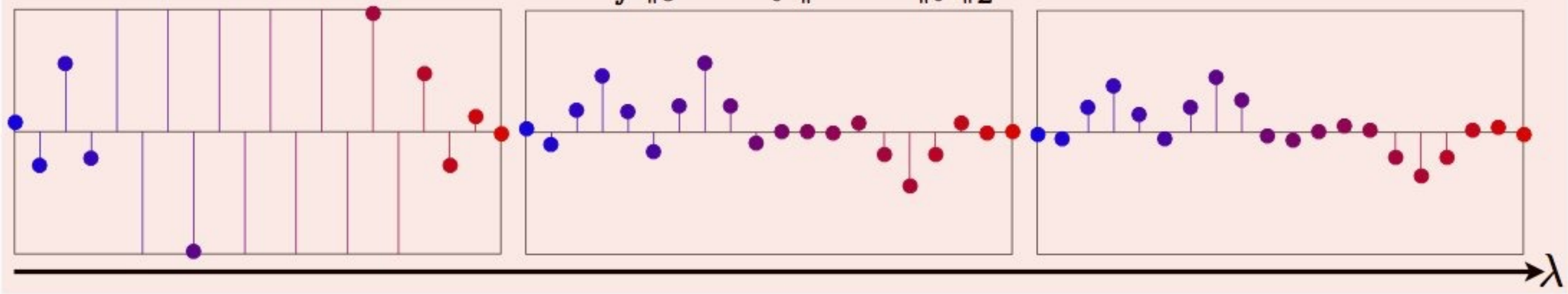


Observations y

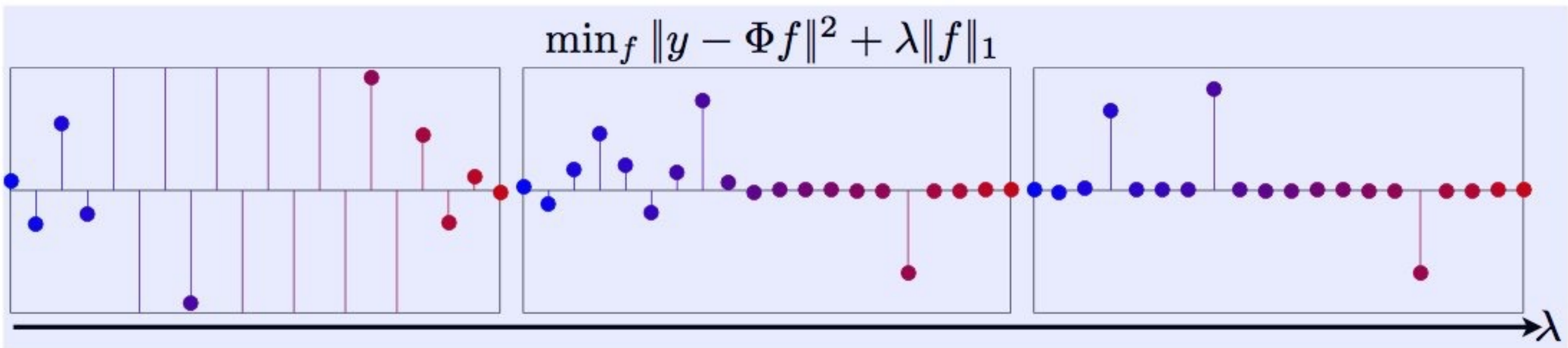


Columns of Φ

$$\min_f \|y - \Phi f\|^2 + \lambda \|f\|_2^2$$



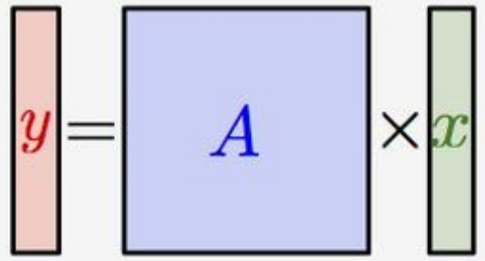
$$\min_f \|y - \Phi f\|^2 + \lambda \|f\|_1$$



L2 vs L1 regularization for inverse problems / correlated designs.

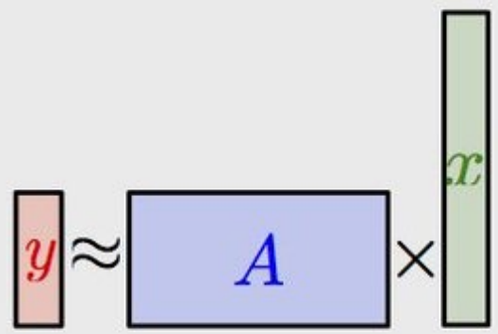
Solving $y \underset{\approx}{=} A\,x \in \mathbb{R}^m$ $\qquad A \in \mathbb{R}^{m \times n}$

Determined $(m = n)$: $\qquad x = A^{-1} y$



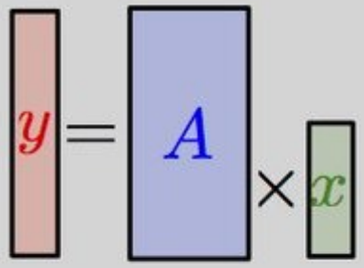
Over-determined $(m > n)$: $\qquad \min_x \|Ax - y\|^2$

$$x = (A^\top A)^{-1} A^\top y \overset{\text{def.}}{=} A^+ y$$



Under-determined $(m < n)$: $\qquad \min_x \{\|x\| \;;\; Ax = y\}$

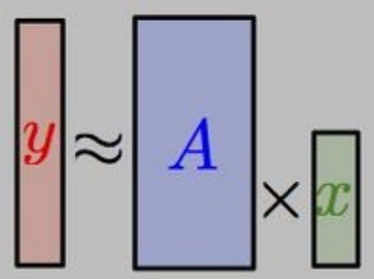
$$x = A^\top (AA^\top)^{-1} y \overset{\text{def.}}{=} A^+ y$$



A ill-posed and/or noise: $\qquad \min_x \|Ax - y\|^2 + \lambda \|x\|^2$

$$x = (A^\top A + \lambda \mathrm{Id}_n)^{-1} A^\top y \overset{\lambda \to 0}{\longrightarrow} A^+ y$$

$$= A^\top (AA^\top + \lambda \mathrm{Id}_m)^{-1} y \quad \text{(Woodbury identity)}$$



Woodbury formula: so simple, so useful ... https://en.wikipedia.org/wiki/Woodbury_matrix_identity …

Automatic Differentiation

How to compute $\nabla \ell_{x,y}(\theta)$? $\quad \ell_{x,y}(\theta) \stackrel{\text{def.}}{=} L(f(x,\theta), y)$

Chain rule: $\nabla \ell_{x,y}(\theta) = [\partial f(x,\theta)]^\top (\nabla L(f(x,\theta), y))$

Linear $f(x,\theta) = \theta \times x$: $\partial f(x,\theta) = \theta$.

Non-linear $f(x,\theta)$: painful ... but $\ell_{x,y}$ it is just a computer program.

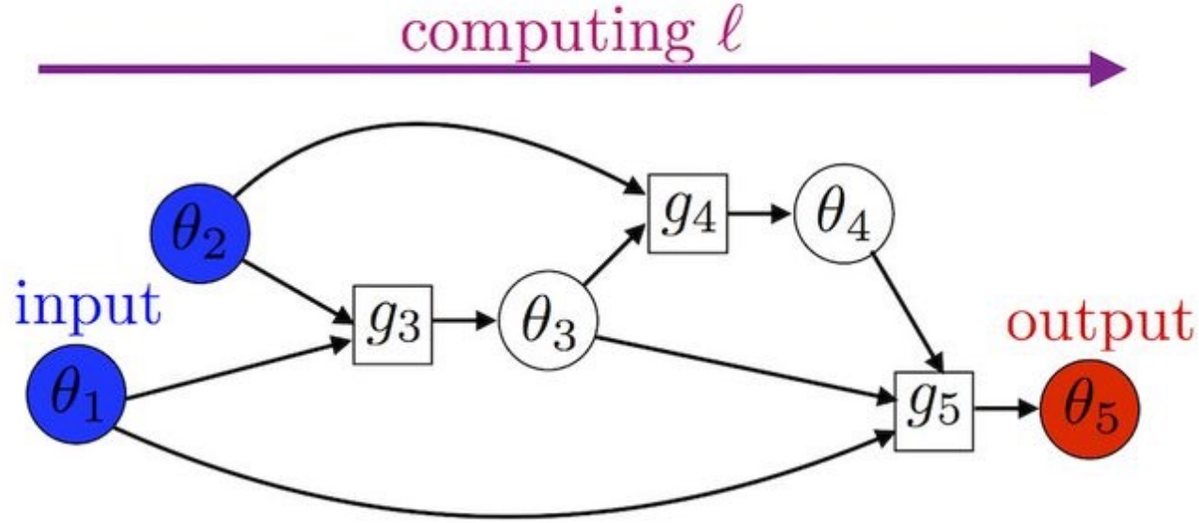
Computer program $\Leftrightarrow$ directed acyclic graph $\Leftrightarrow$ linear ordering of nodes $(\theta_r)_r$

forward

```
function ℓ(θ_1, ..., θ_M)
  for r = M+1, ..., R
    θ_r = g_r(θ_Parents(r))
  return θ_R
```

backward

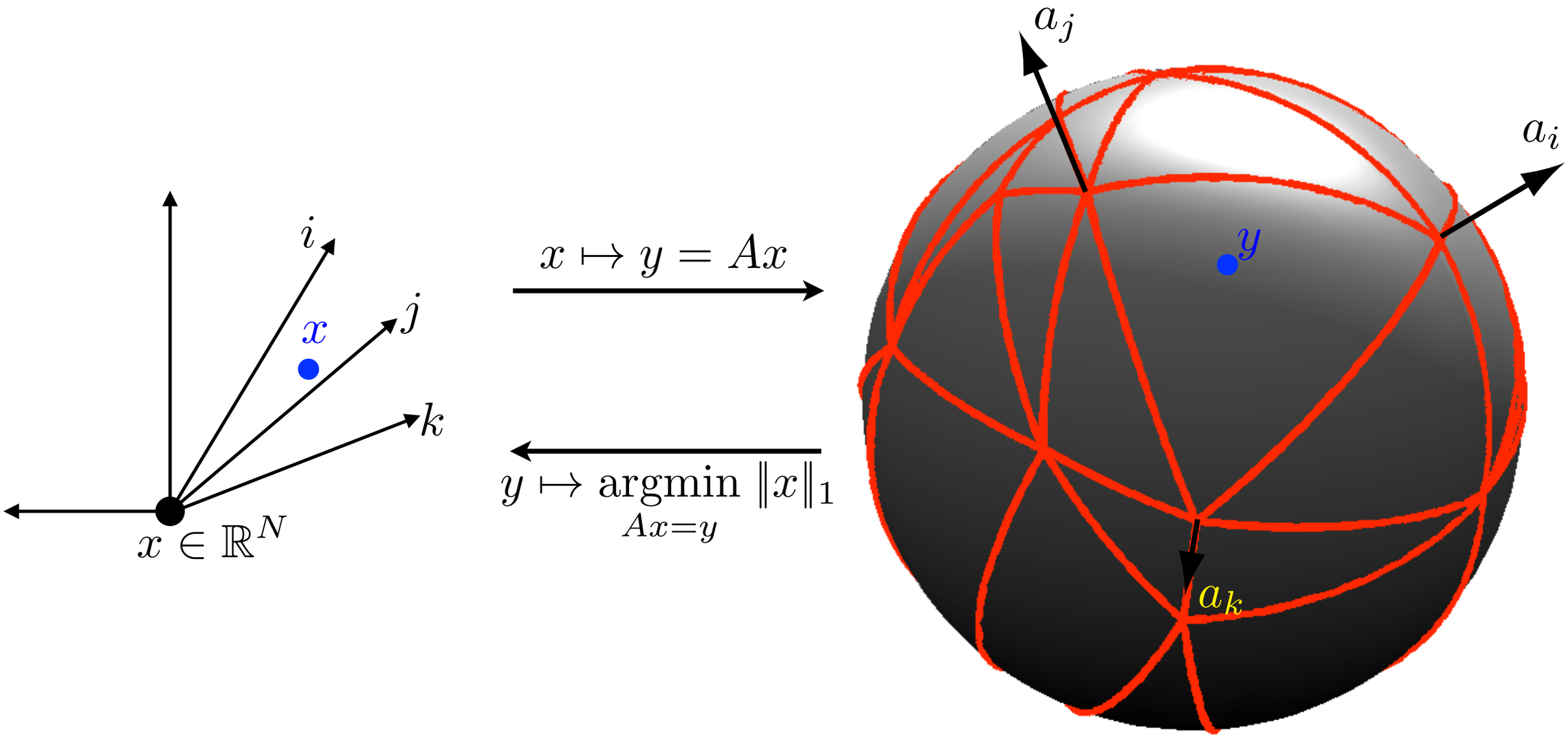
```
function ∇ℓ(θ_1, ..., θ_M)
  ∇_R ℓ = 1
  for r = R−1, ..., 1
    ∇_r ℓ = Σ_{s∈Child(r)} ∂_r g_s(θ) ∇_s ℓ
  return (∇_1 ℓ, ..., ∇_M ℓ)
```



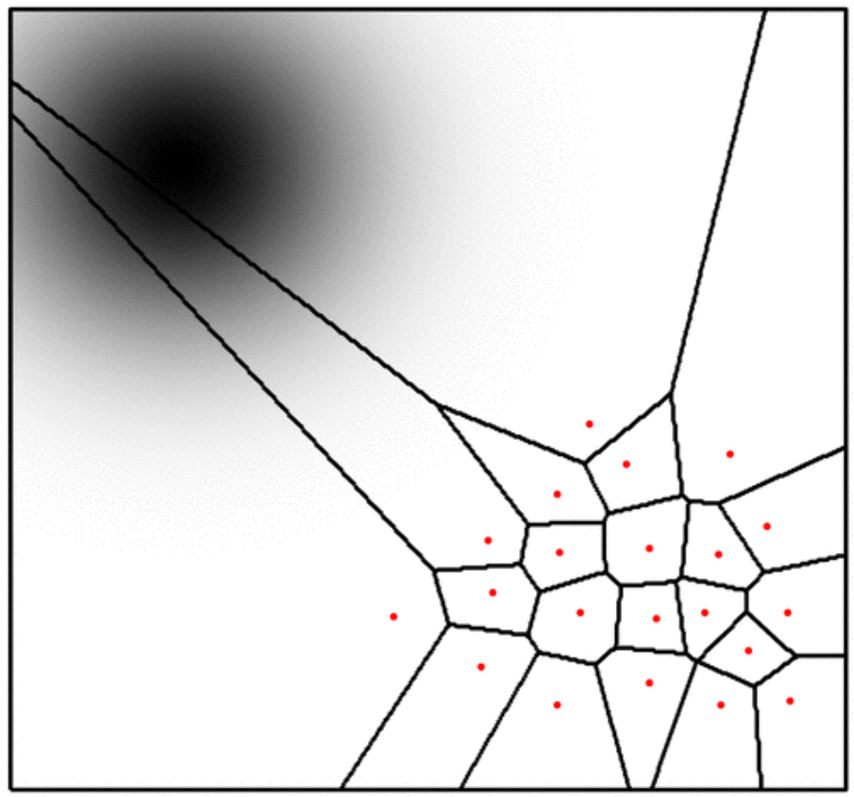
$$\text{“}\frac{\partial \ell}{\partial \theta_r} = \sum_{s \in \text{Child}(r)} \frac{\partial \ell}{\partial \theta_s} \frac{\partial \theta_s}{\partial \theta_r}\text{”}$$

$\nabla_r \ell(\theta)$ $\quad$ $\nabla_s \ell(\theta)$ $\quad$ $\partial_r g_s(\theta)$

Reverse mode auto-diff in 1 slide. So simple. So useful.



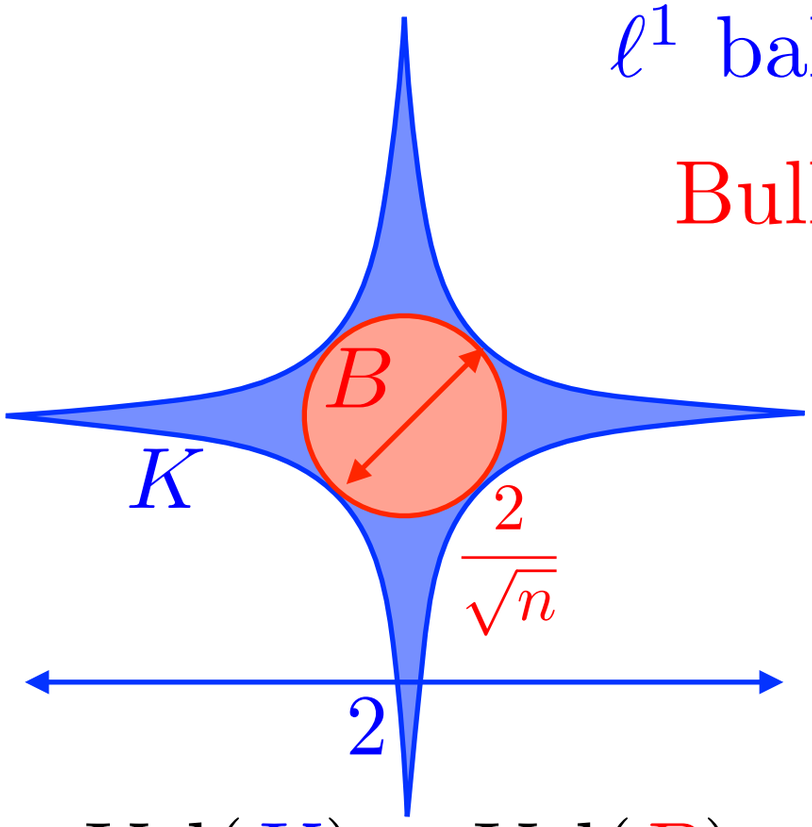
Signs of sparse vectors recovered by l1 minimization index a spherical Delaunay triangulation of the observation space. More on this in my lecture notes. https://mathematical-tours.github.io/book-sources/chapters-pdf/sparse-theory.pdf …



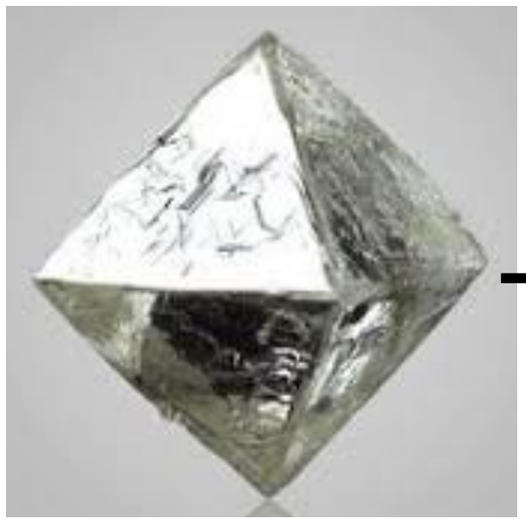
Lloyd algorithm (k-means) with non-uniform sampling density. https://en.wikipedia.org/wiki/Lloyd%27s_algorithm …

ℓ^1 ball: $K \stackrel{\text{def.}}{=} \left\{x \in \mathbb{R}^n \; ; \; \sum_{i=1}^n |x_i| \leqslant 1\right\}$

Bulk: $B \stackrel{\text{def.}}{=} \left\{x \in \mathbb{R}^n \; ; \; \sum_{i=1}^n |x_i|^2 \leqslant n^{-1}\right\}$



$\mathrm{Vol}(K) \sim \mathrm{Vol}(B)$



small n

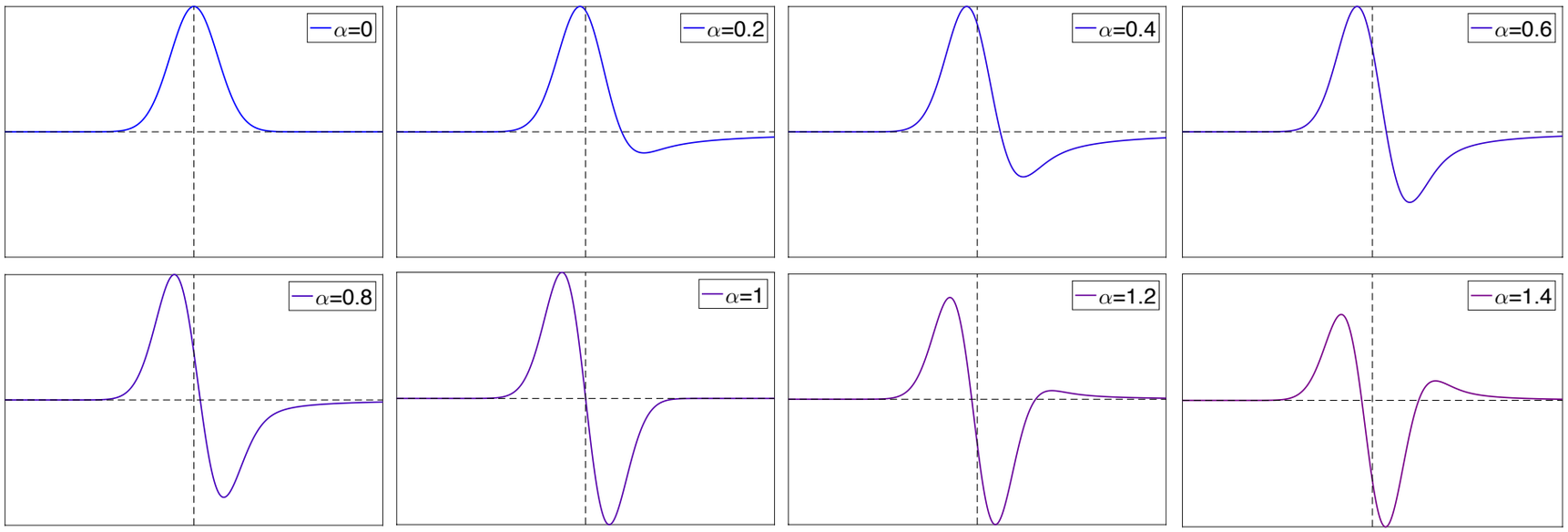
large n

Vitali Milman's insight: convex sets in high dimension look like hedgehogs http://www.math.tau.ac.il/~milman/files/survey4.ps

Signal and Image Processing

Fourier transform: $\mathcal{F}(f)(\omega) \overset{\text{def.}}{=} \int_{\mathbb{R}} f(x) e^{-\mathrm{i}\omega x} \mathrm{d}x$

Fractional derivative: $\mathcal{F}(f^{(\alpha)}) \overset{\text{def.}}{=} (\mathrm{i}\omega)^{\alpha} \mathcal{F}(f)(\omega)$



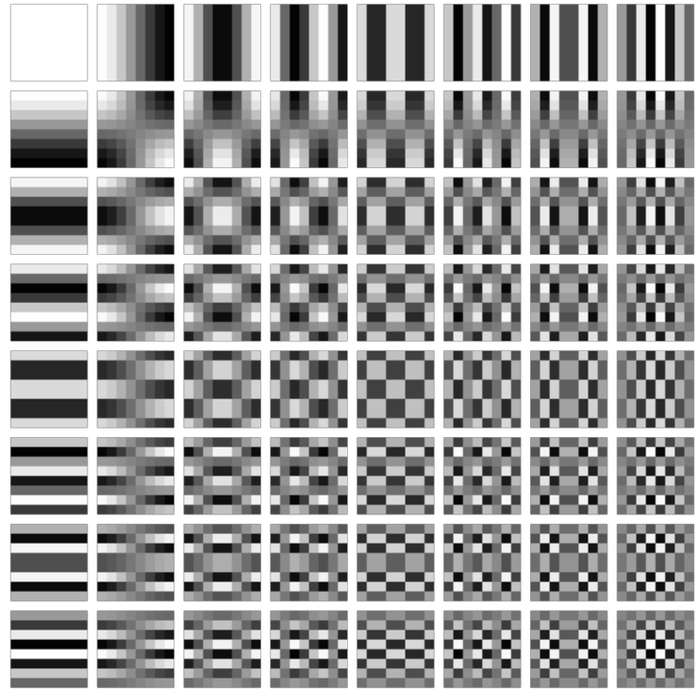
Fractional derivative of a Gaussian using the Fourier transform. Code provided. https://en.wikipedia.org/wiki/Fractional_calculus … https://goo.gl/9CJMmD

DCT-1: $\cos jk\frac{\pi}{N-1}$ (divide by $\sqrt{2}$ when j or k is 0 or $N-1$)

DCT-2: $\cos\left(j+\frac{1}{2}\right)k\frac{\pi}{N}$ (divide by $\sqrt{2}$ when $k=0$)

DCT-3: $\cos j\left(k+\frac{1}{2}\right)\frac{\pi}{N}$ (divide by $\sqrt{2}$ when $j=0$)

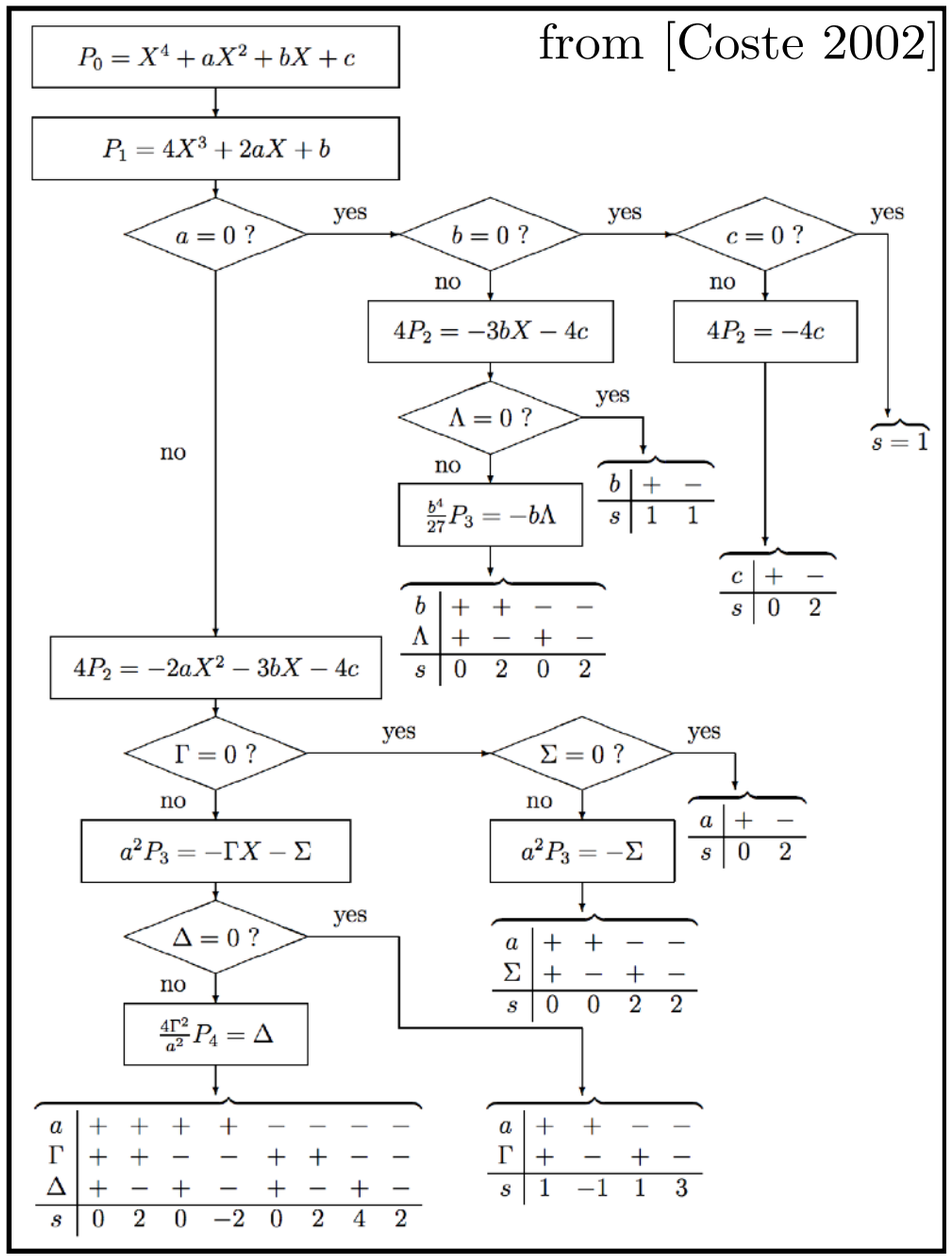
DCT-4: $\cos\left(j+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{N}$



The discrete case has a new level of variety and complexity, often appearing in the boundary conditions [G. Strang - SIAM review, 1999]

Discrete Cosine Transforms are eigenvectors of Laplacians with different boundary cond. Wonderful paper by Strang. http://www-math.mit.edu/~gs/papers/dct.pdf …

Ongoing



algebraic set

$$\mathcal{X} \stackrel{\text{def.}}{=} \{(a,b,c,X)\ ;\ X^4 + aX^2 + bX + c = 0\}$$

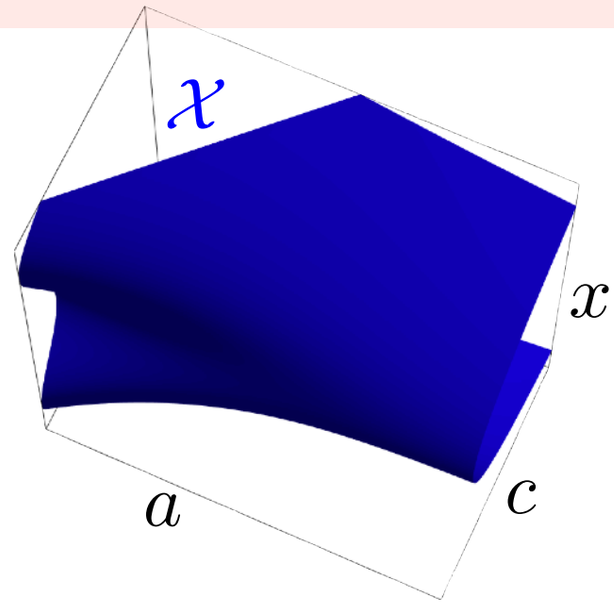
projection

$$(a,b,c,X) \mapsto (a,b,c)$$

$$\{(a,b,c)\ ;\ \exists X \in \mathbb{R}, X^4 + aX^2 + bX + c = 0\}$$

semi-algebraic set

$b = 0$

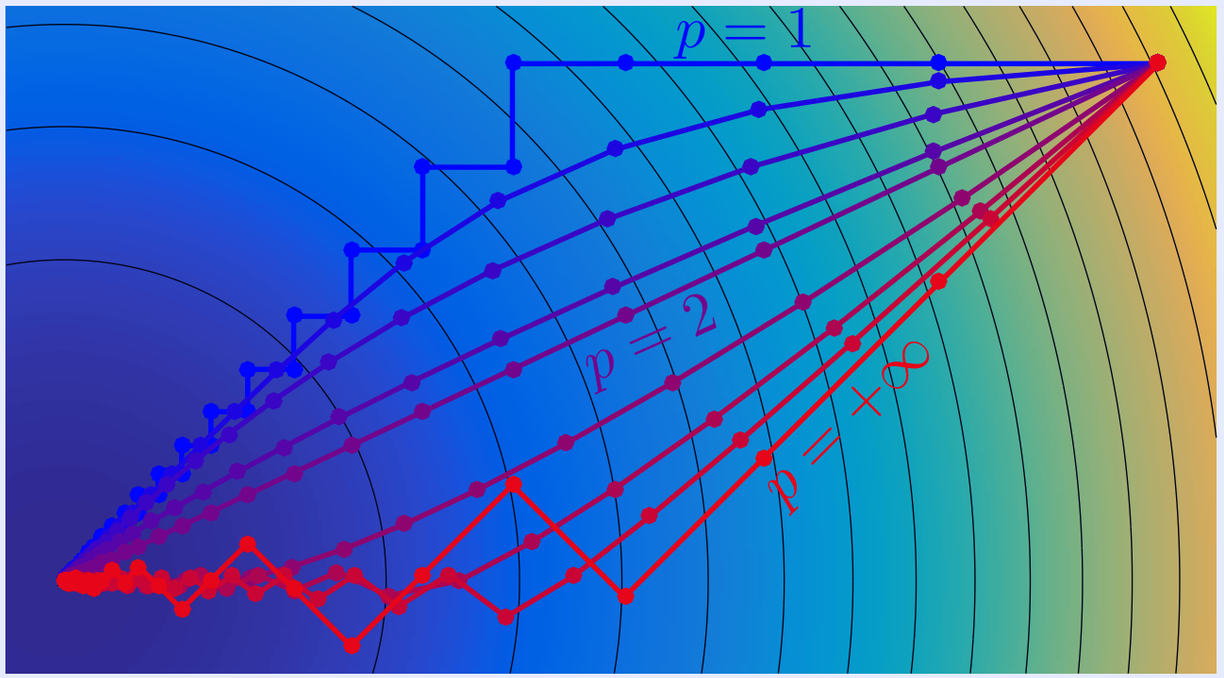


Tarski–Seidenberg in action: the set of polynomials having a real root is semi-algebraic. Coste provides the block diagram for degree 4 in his notes. Amazing.
http://gcomte.perso.math.cnrs.fr/M2/CosteIntroToSemialGeo.pdf

Metric space $(\mathcal{X}, d)$, minimize $F(x)$ on $\mathcal{X}$.

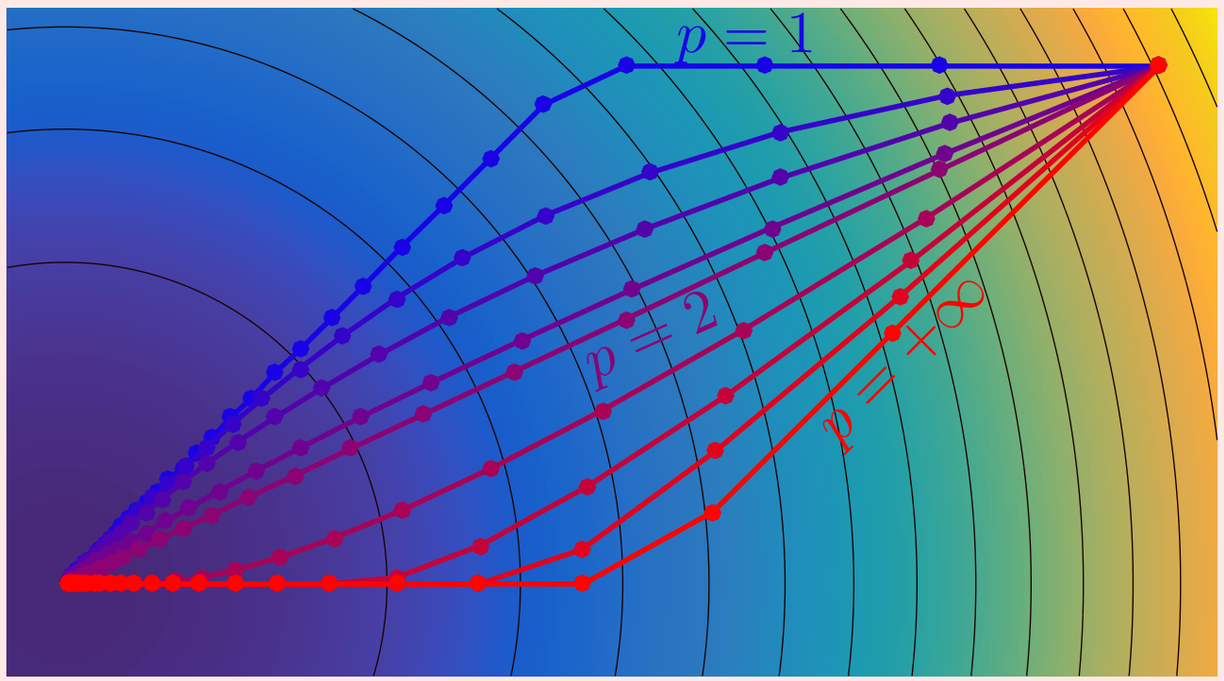
Explicit

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} d(x_k, x)^2 + \tau \langle \nabla F(x_k),\, x \rangle$$



Implicit

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} d(x_k, x)^2 + \tau F(x)$$



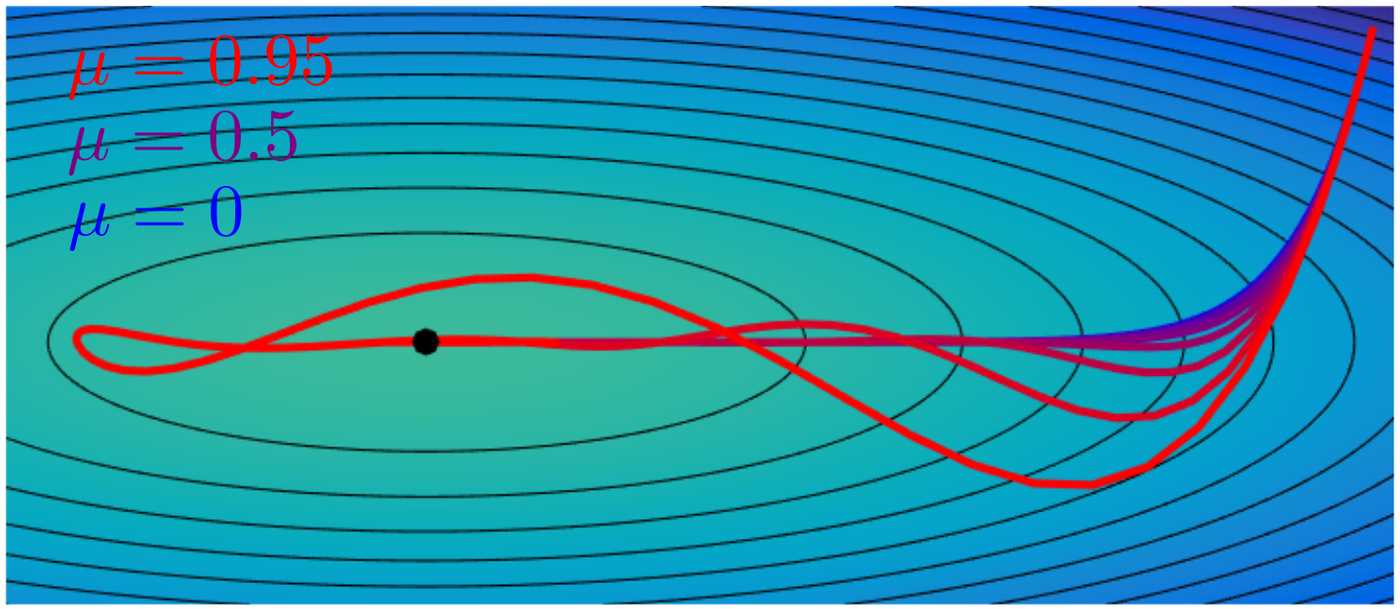
$$F(x) = \|x\|^2 \text{ on } (\mathcal{X} = \mathbb{R}^2, \|\cdot\|_p)$$

Comparison of explicit stepping (unstable) vs implicit stepping (stable) for gradient flow on metric space. Note how L1 flow is equivalent to a coordinate descent, and L2 is the usual gradient descent.

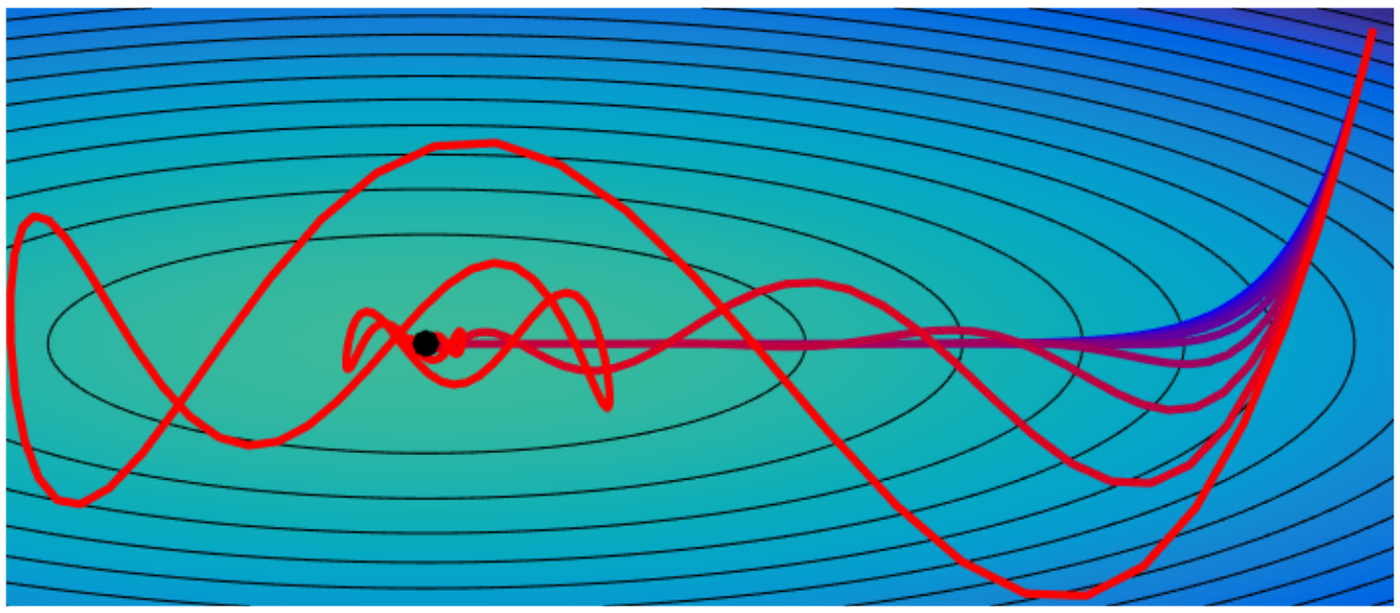
$$x_{k+1} = x_k + p_k$$

$$p_{k+1} = \mu p_k - \tau \begin{cases} \nabla f(x_k) & \text{Polyak} \\ \nabla f(x_k + \mu p_k) & \text{Nesterov} \end{cases}$$

Yurii
Nesterov



Boris
Polyak



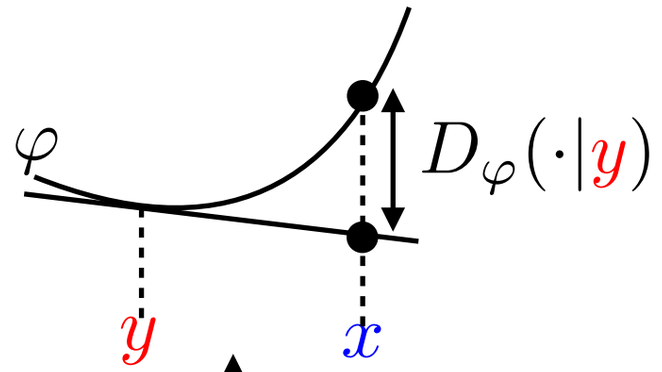
Boris Polyak's heavy ball and Yurii Nesterov's descent use momentum to accelerate but oscillate.

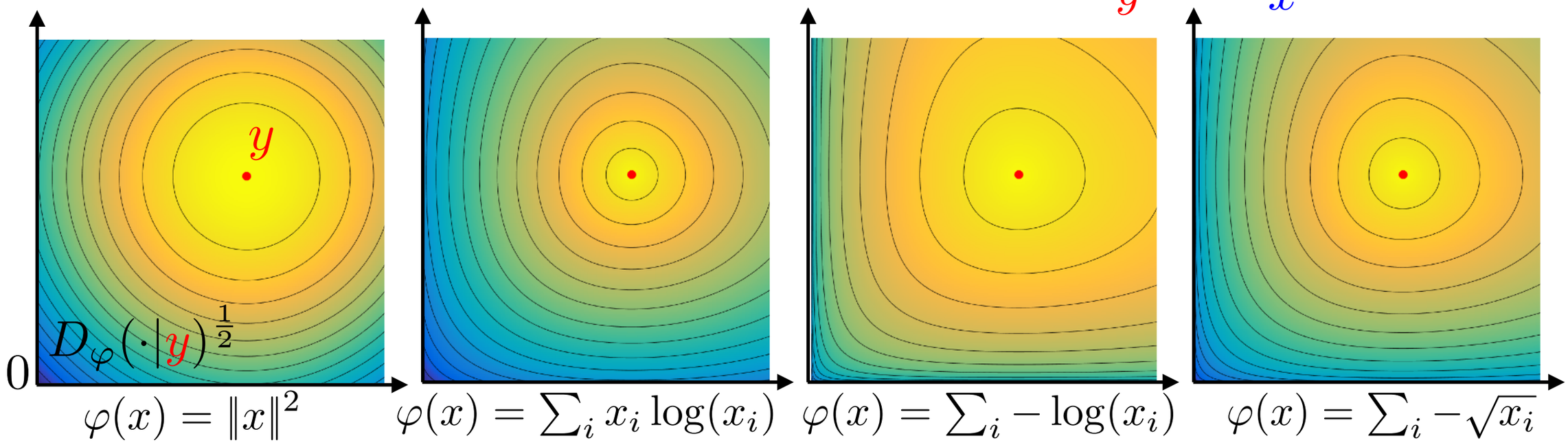
http://lab7.ipu.ru/eng/people/polyak.html

https://fr.wikipedia.org/wiki/Yurii_Nesterov

Bregman divergence:

$$D_\varphi(x|y) \overset{\text{def.}}{=} \varphi(x) - \varphi(y) - \langle x - y, \nabla\varphi(y)\rangle$$





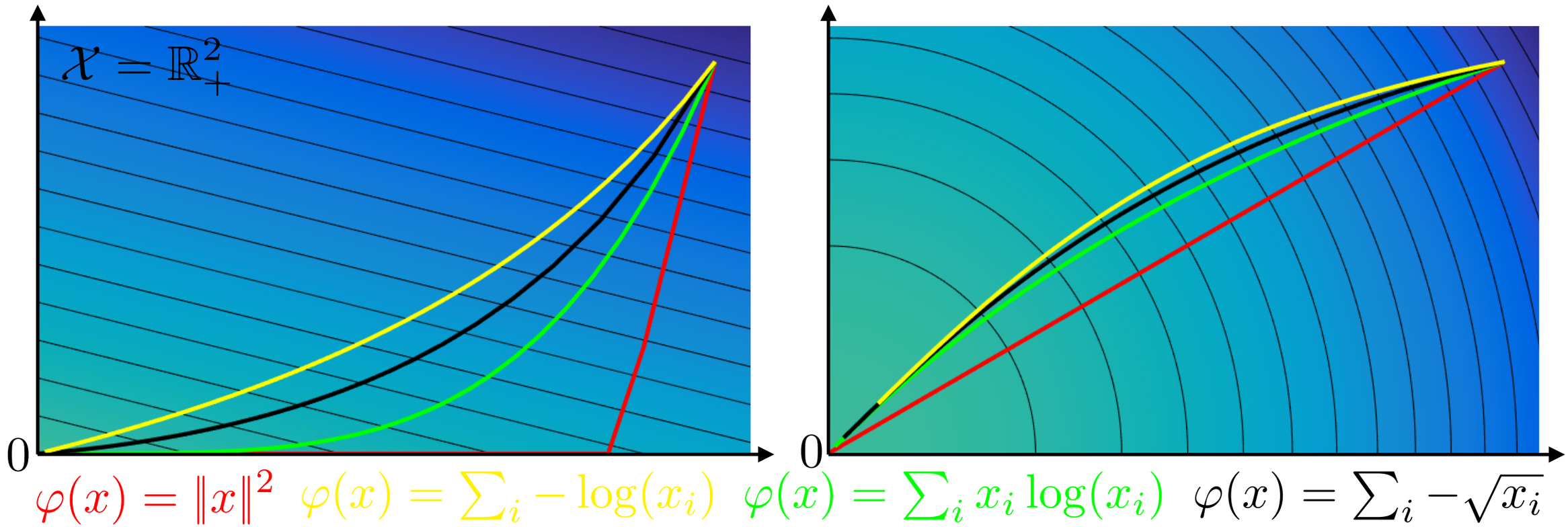
Bregman divergences are convex distance-like functionals which are locally Euclidean. Any reasonable algorithm handling Euclidean distances generalizes to Bregman geometries.

https://en.wikipedia.org/wiki/Bregman_divergence

$$\text{Bregman divergence:} \quad D_\varphi(x|y) \overset{\text{def.}}{=} \varphi(x) - \varphi(y) - \langle x - y,\, \nabla\varphi(y)\rangle$$

$$\text{Mirror descent:} \quad x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} D_\varphi(x|x_k) + \tau\langle \nabla f(x_k),\, x\rangle$$

$$= (\nabla\varphi)^{-1}\left(\nabla\varphi(x_k) - \tau\nabla f(x_k)\right)$$



Mirror descent generalizes gradient descent using Bregman geometries.

https://blogs.princeton.edu/imabandit/2013/04/16/orf523-mirror-descent-part-iii/