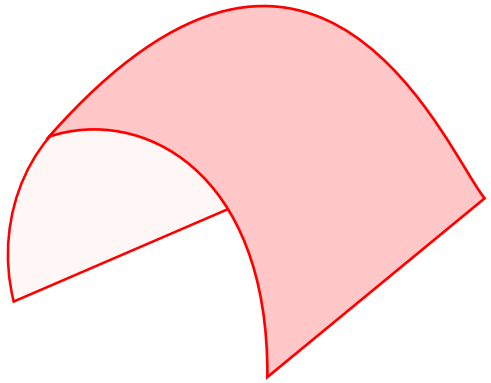
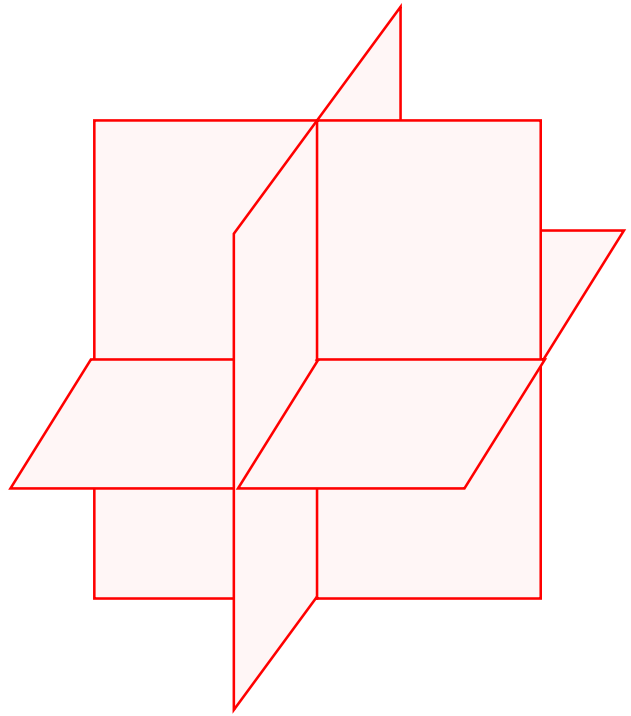
Don’t Relax!

The unreasonable effectiveness of non-convex optimization

Purushottam Kar

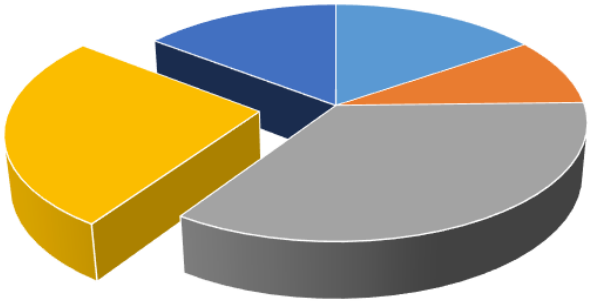
IIT KANPUR



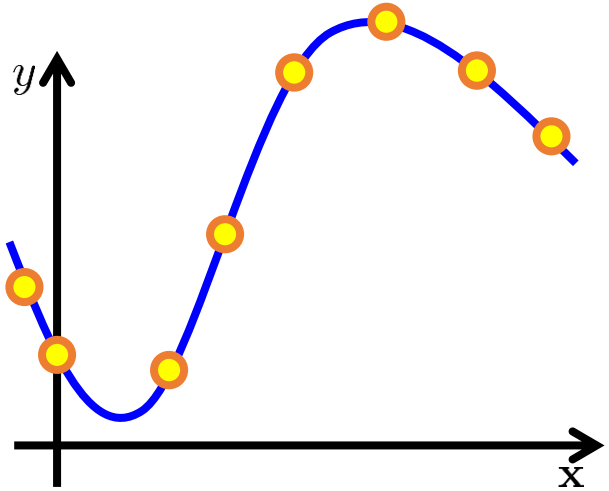


Optimization is Everywhere!

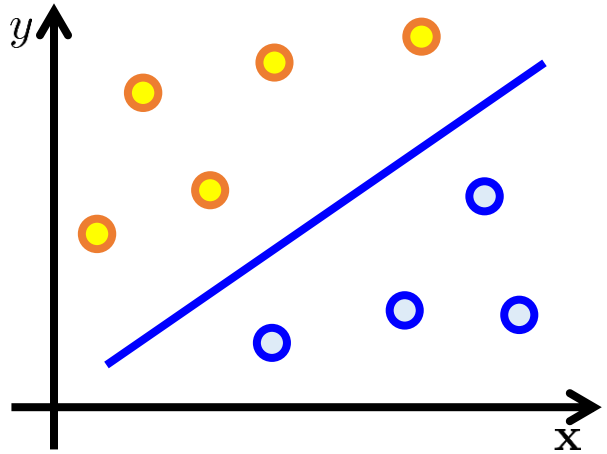
Applications



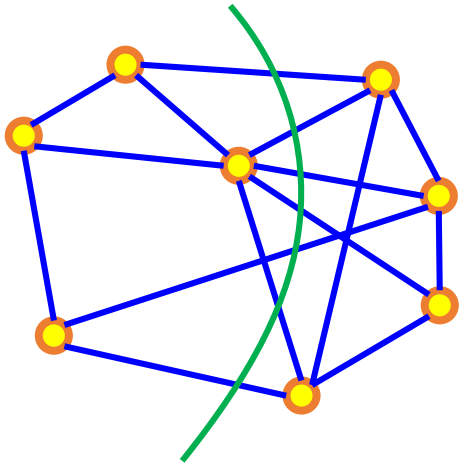
Resource Allocation



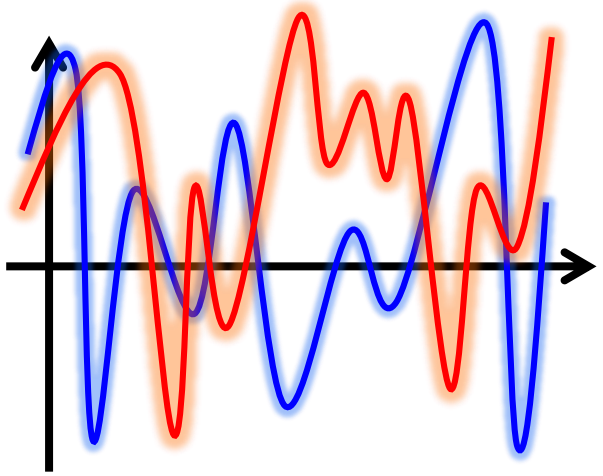
Regression



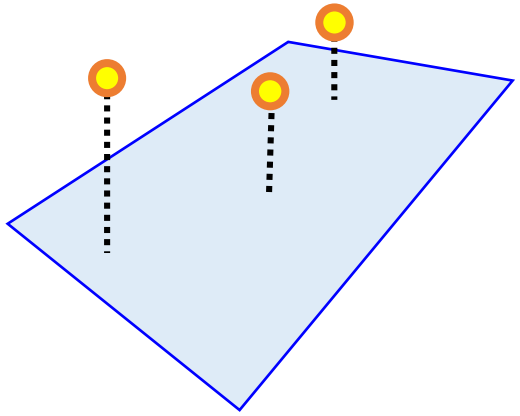
Classification



Clustering/Partitioning



Signal Processing



Dimensionality Reduction

Convex Optimization

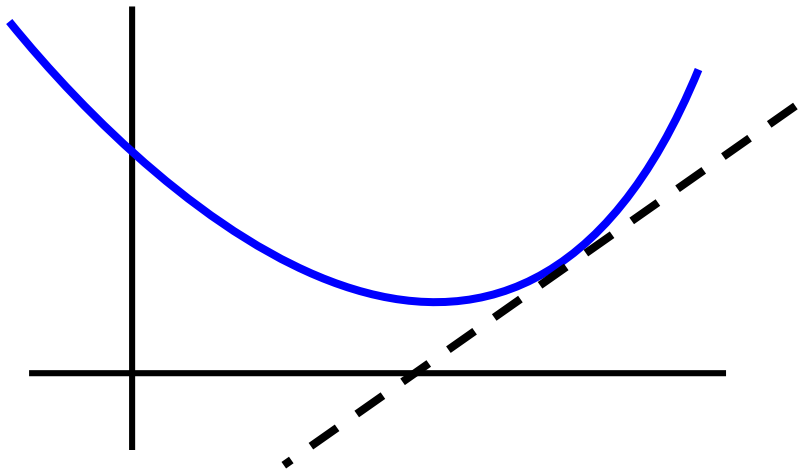
$$\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x})$$

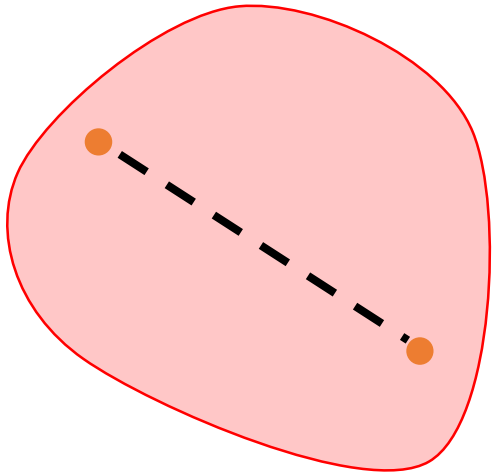
$$f : \mathbb{R}^d \rightarrow \mathbb{R}$$

Convex function

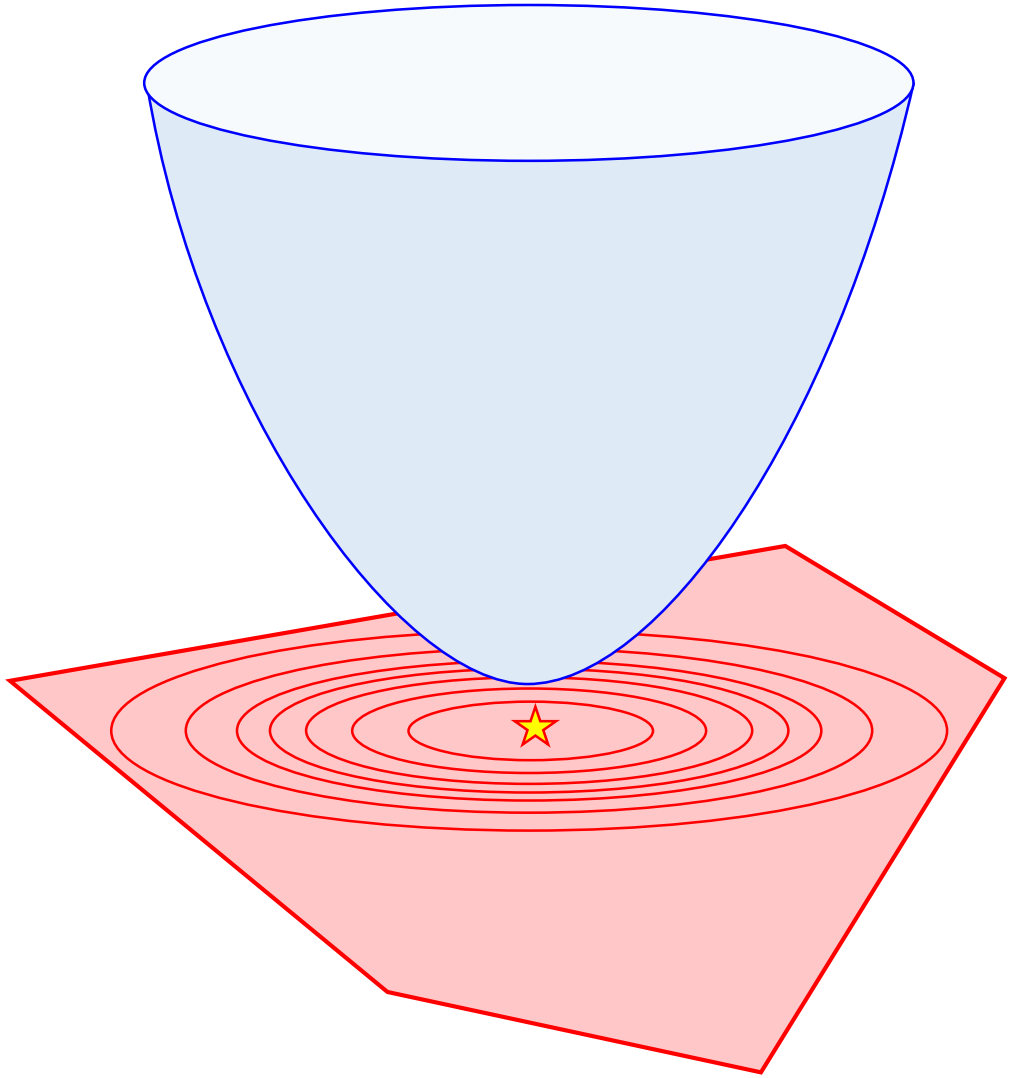
$$\mathcal{C} \subseteq \mathbb{R}^d$$

Convex set

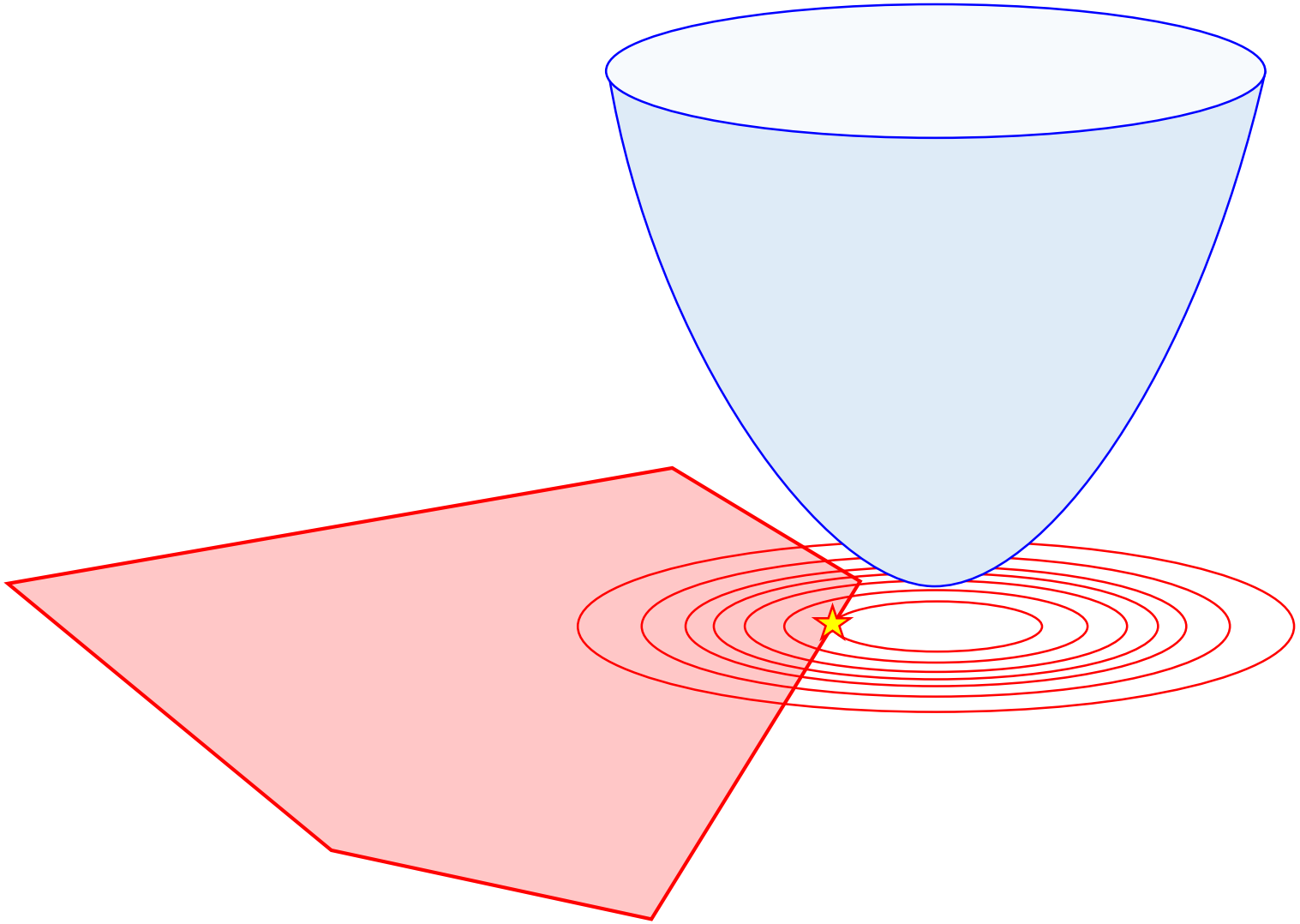




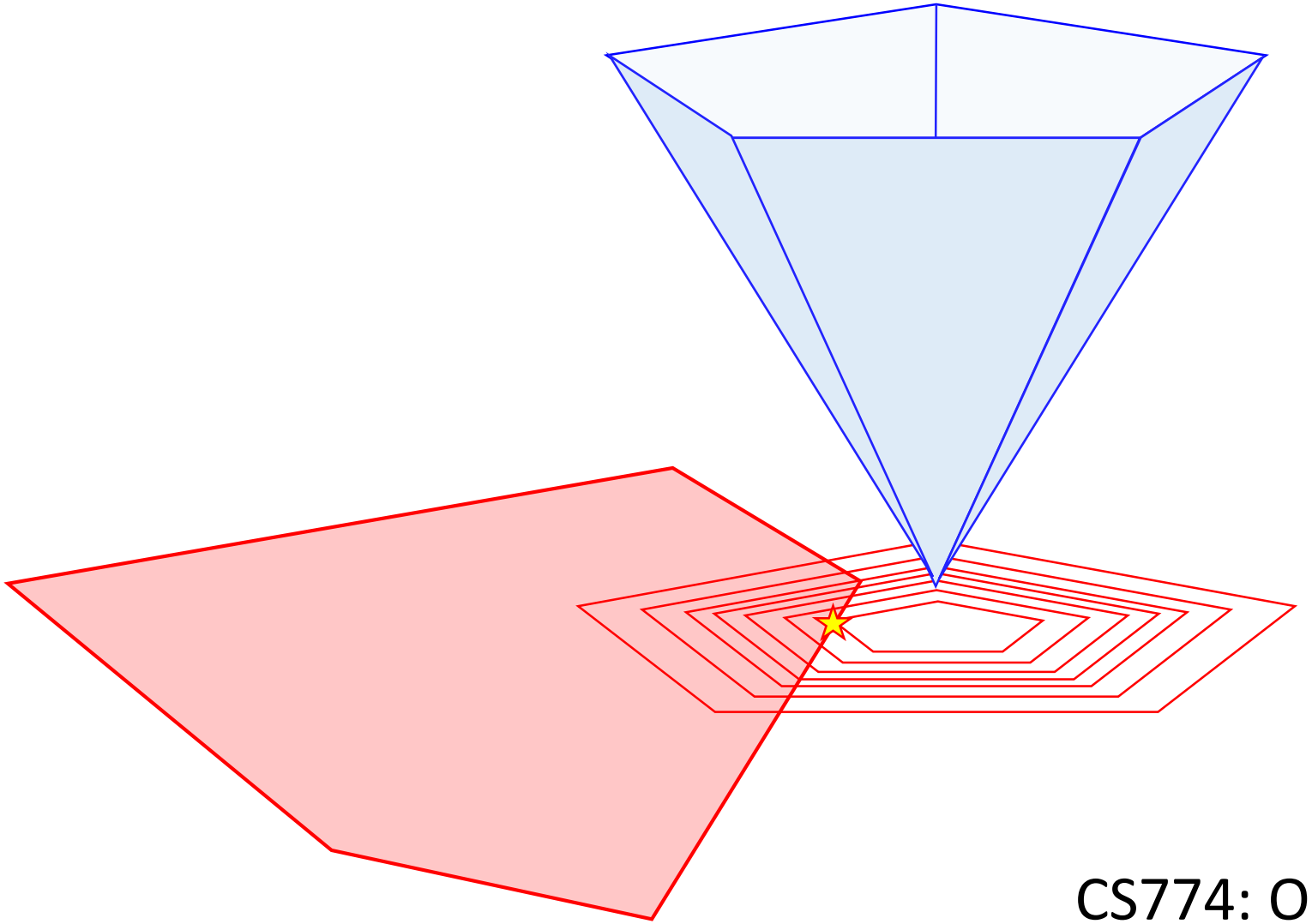
A Cartoon View of Optimization



A Cartoon View of Optimization

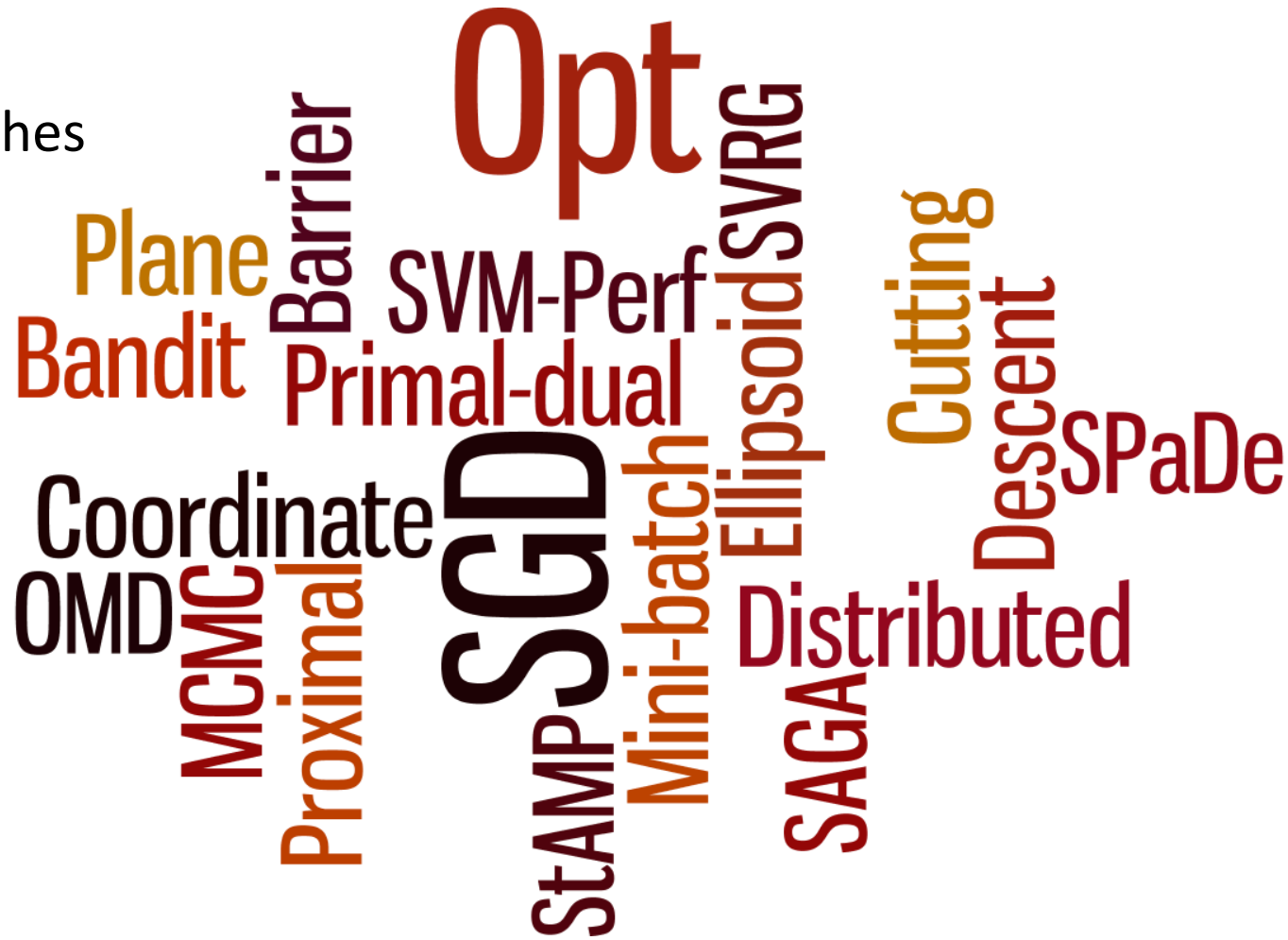


A Cartoon View of Optimization



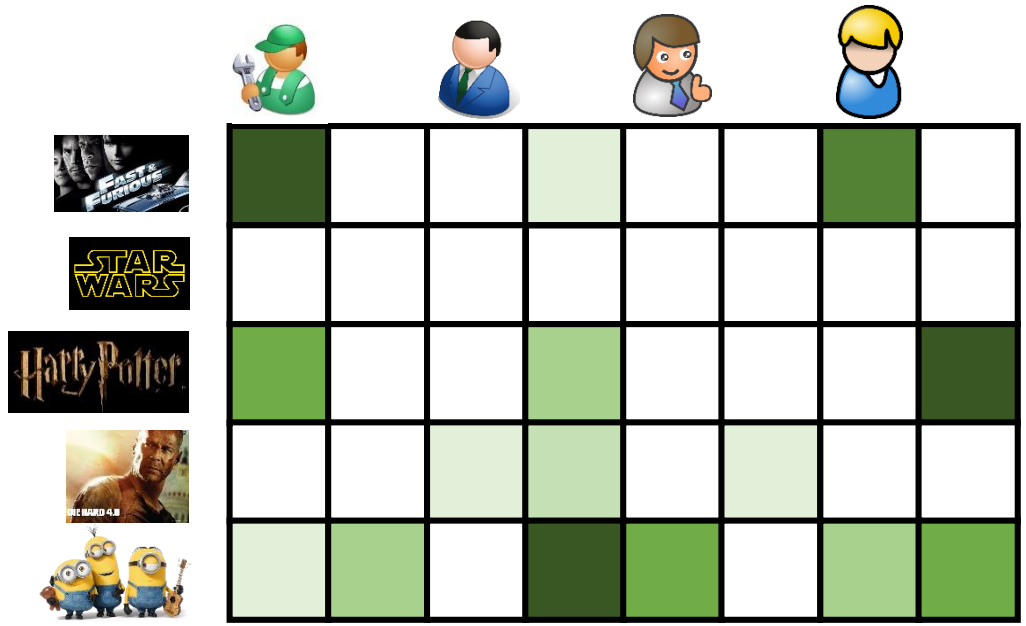
Techniques

- Projected (Sub)gradient Methods
  - Stochastic, mini-batch variants
  - Primal, dual, primal-dual approaches
  - Coordinate update techniques
- Interior Point Methods
  - Barrier methods
  - Annealing methods
- Other Methods
  - Cutting plane methods
  - Accelerated routines
  - Proximal methods
  - Distributed optimization
  - Derivative-free optimization



Non-convex Optimization?

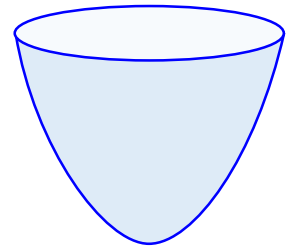
Recommender Systems



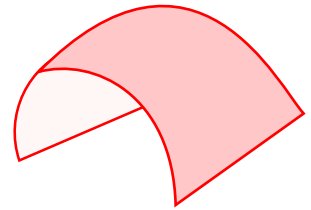
$$\min_{L \in \mathbb{R}^{m \times n}} \|X_\Omega - L_\Omega\|_F^2$$

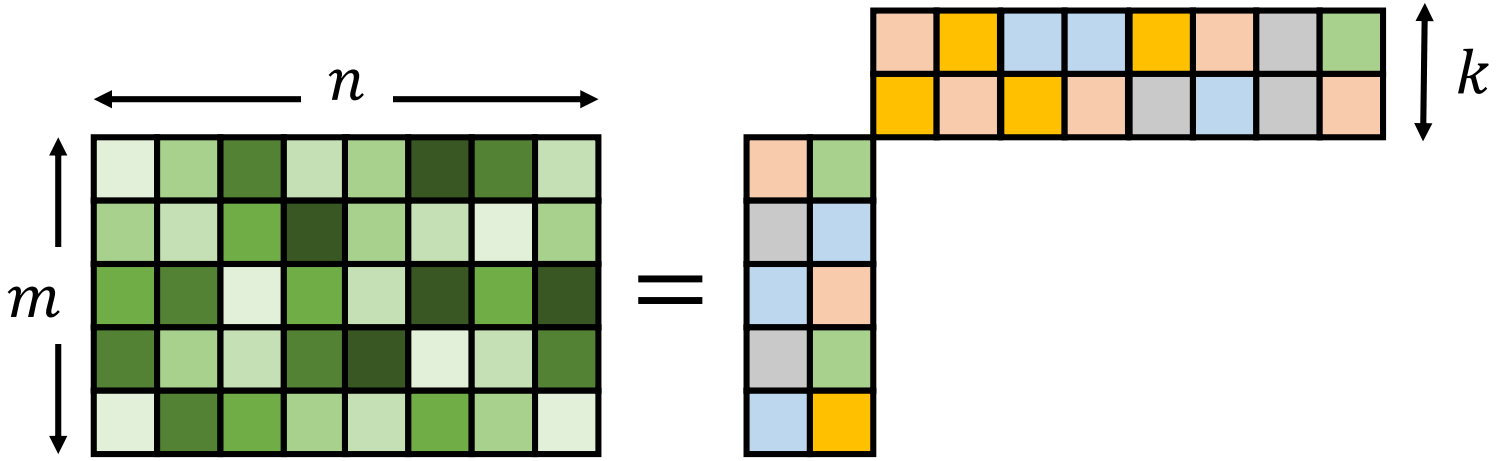
$$s.t.\ \mathrm{rank}(L) \leq k$$

f



$\mathcal{C}$

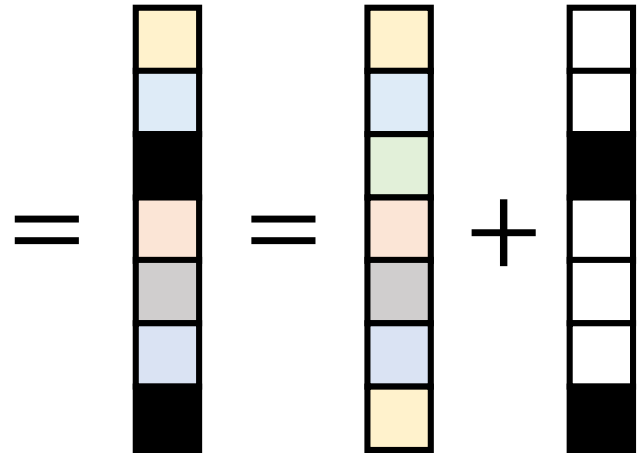




clker.com, clipartpanda.com, nj-clucker.com, hotelgardenmoshir.com

Large Scale Surveillance

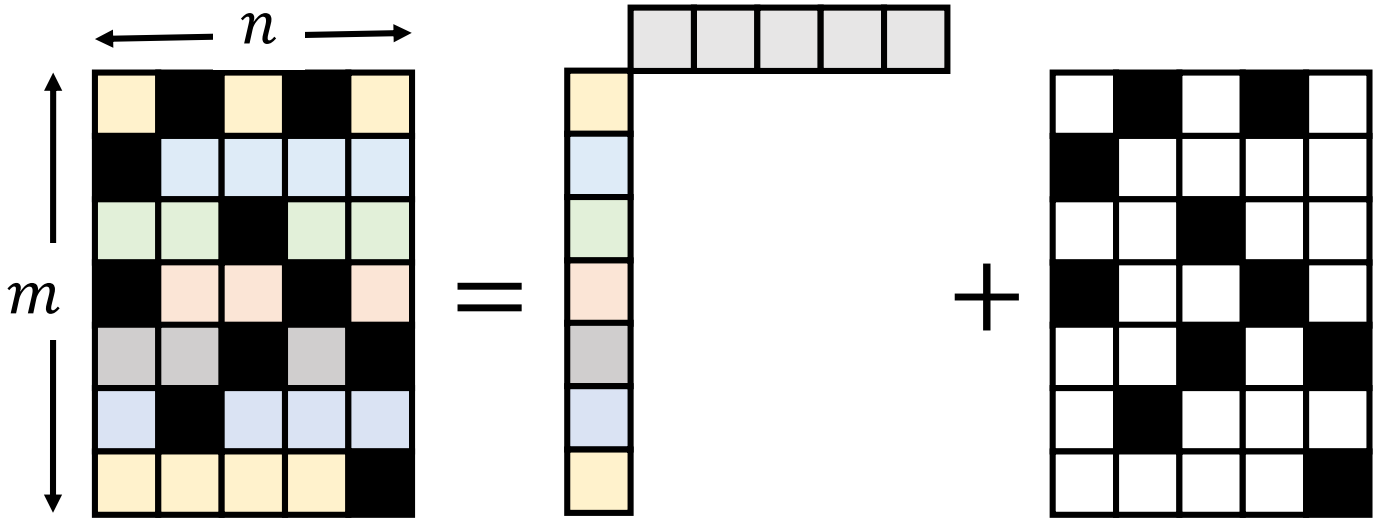
- Foreground-background separation



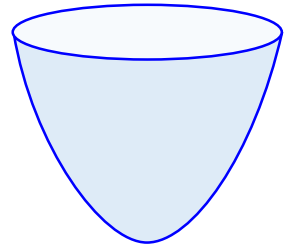
$$\min_{L,S} \|X - (L+S)\|_F^2$$

$$s.t.\ \mathrm{rank}(L) \leq k$$

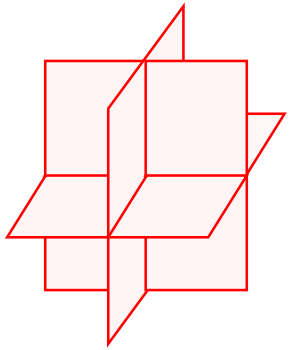
$$\|S\|_0 \leq s \ll mn$$

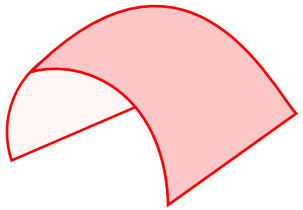


f



$\mathcal{C}$





X-ray Crystallography

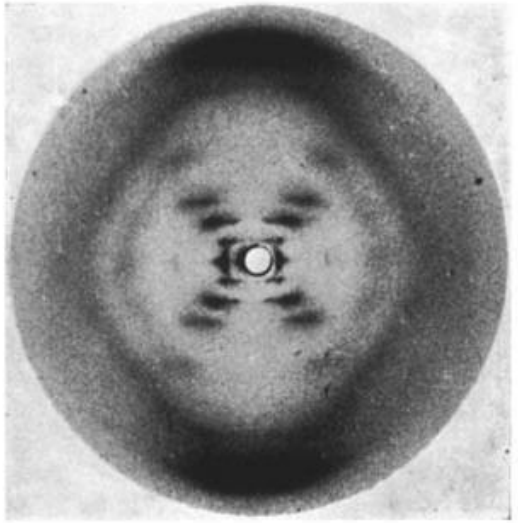
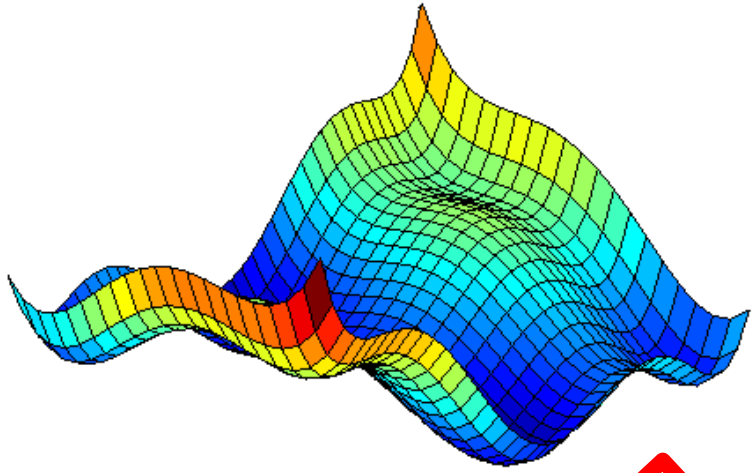
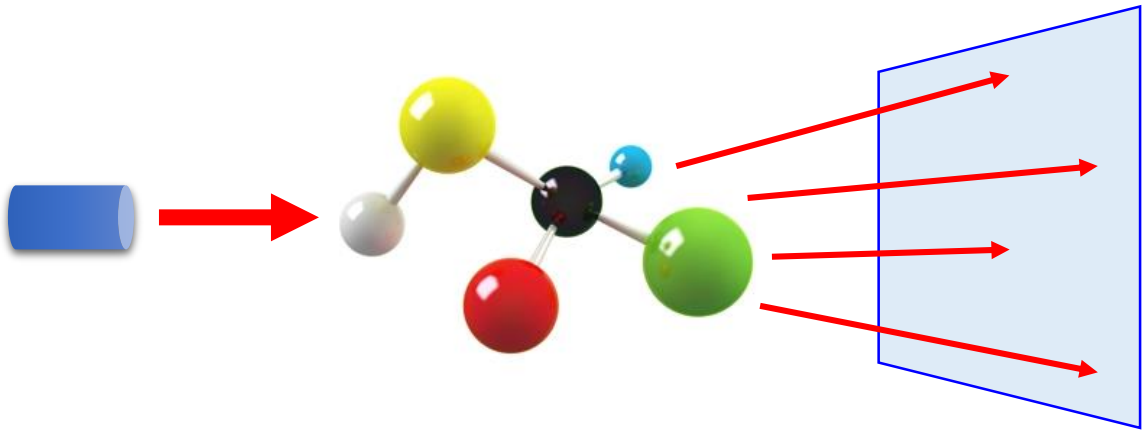


Photo 51 that led to the discovery of the helical structure of DNA

$$\min_{\mathbf{w}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i^2 - (\mathbf{x}_i^\top \mathbf{w})^2)^2$$

f





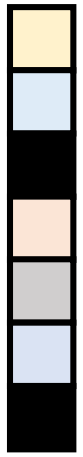
$$y_i = \left|\mathbf{x}_i^\top \mathbf{w}^*\right|$$

Phase Recovery!

tes.com, marioantonelli.it, wikipedia.com

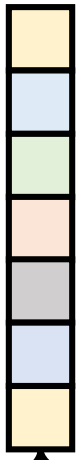
Image Denoising and Robust Face Recognition

$$=$$



$$\mathbf{y}$$

$$=$$

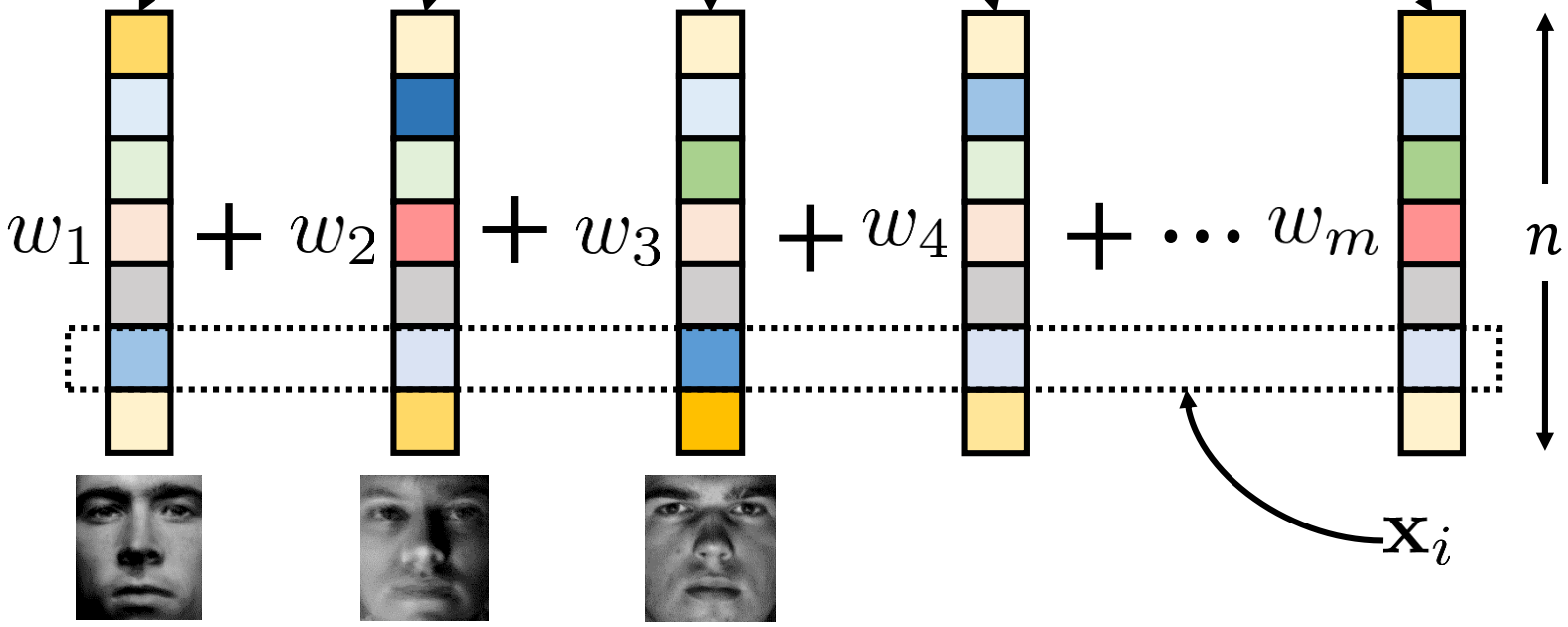


$$+$$

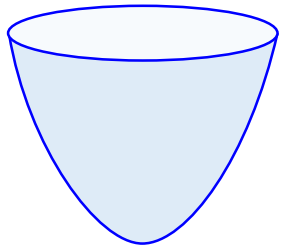
$$\mathbf{b}$$

$$\min_{\mathbf{w}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w} - b_i)^2$$

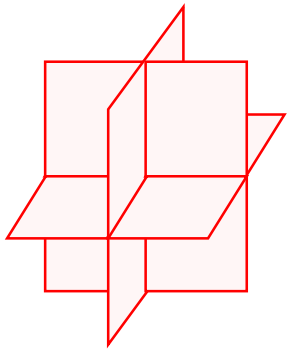
$$s.t.\ \|\mathbf{b}\|_0 \le k \ll n$$



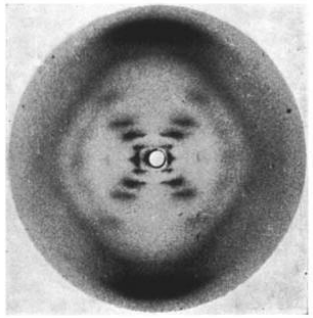
f



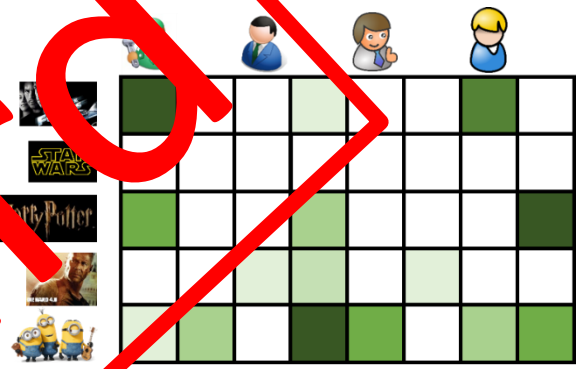
$\mathcal{C}$



Non Convex Optimization

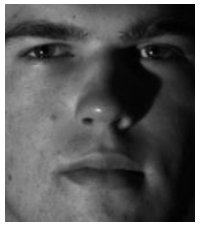


Phase Retrieval



Matrix Completion

NP-hard



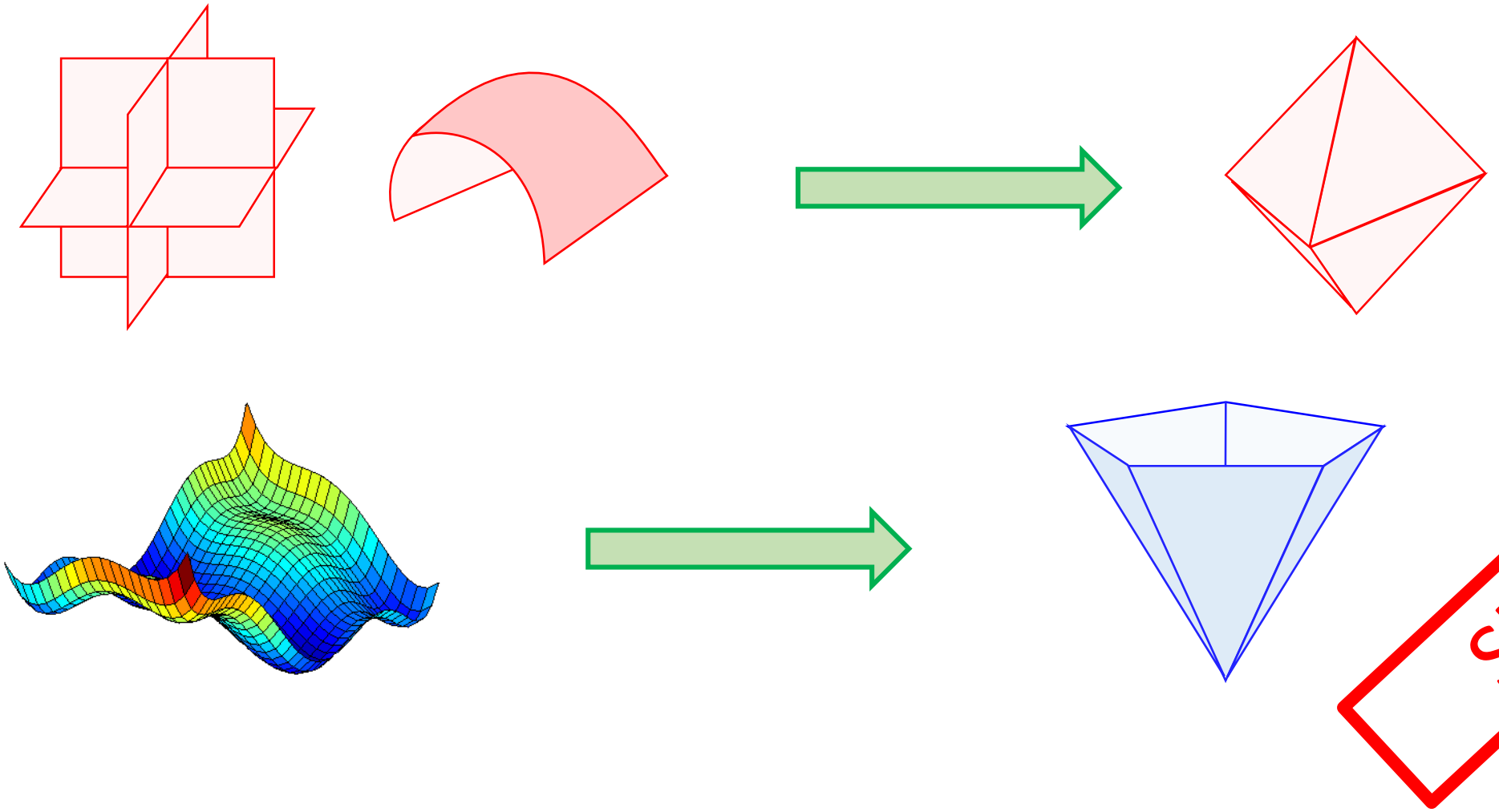
Robust Regression

Robust PCA

Non-convex Optimization: A Brief Introduction

Relaxation-based Techniques

- “Convexify” the feasible set and the objective



Alternating Minimization

$$\begin{aligned} \min \ & f(\mathbf{x}, \mathbf{y}) \\ s.t. \ & \mathbf{x} \in \mathcal{C}_1 \\ & \mathbf{y} \in \mathcal{C}_2 \end{aligned}$$

- Initialize $\mathbf{x}^0, \mathbf{y}^0$
- For $t = 1, 2, \ldots$
  - $\mathbf{x}^t = \arg\min_{\mathbf{x} \in \mathcal{C}_1} f(\mathbf{x}, \mathbf{y}^{t-1})$
  - $\mathbf{y}^t = \arg\min_{\mathbf{y} \in \mathcal{C}_2} f(\mathbf{x}^t, \mathbf{y})$

Matrix Completion

$$\min_{L \in \mathcal{M}_k^{m,n}} \|X_\Omega - L_\Omega\|_F^2$$

$$\equiv \min_{\substack{U \in \mathbb{R}^{m \times k} \\ V \in \mathbb{R}^{n \times k}}} \|X_\Omega - (UV^\top)_\Omega\|_F^2$$

Robust PCA

$$\min_{\substack{L \in \mathcal{M}_k^{m,n} \\ S \in \mathcal{B}_0^{m,n}(s)}} \|X - (L + S)\|_F^2$$

Alternating Minimization

http://shabal.in

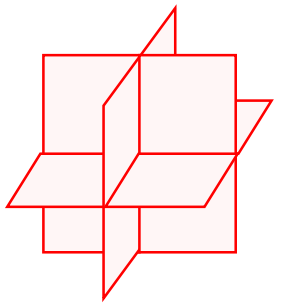
Projected Gradient Descent

$$\min \ f(\mathbf{x})$$
$$s.t. \ \mathbf{x} \in \mathcal{C}$$

- Initialize $\mathbf{x}^0$
- For $t = 1, 2, \ldots$
  - $\mathbf{z}^t = \mathbf{x}^{t-1} - \eta_t \cdot \nabla f(\mathbf{x}^{t-1}))$
  - $\mathbf{x}^t = \Pi_{\mathcal{C}}(\mathbf{z}^t)$

$$\Pi_{\mathcal{C}}(\mathbf{z}) = \arg\min_{\mathbf{x} \in \mathcal{C}} \|\mathbf{z} - \mathbf{x}\|_2^2$$

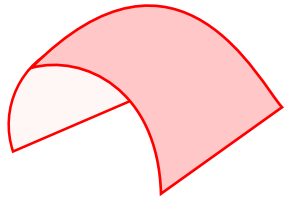
Non-convex Projection



$\mathcal{B}_0^p(s)$

Top s elements by magnitude

$\mathcal{M}_k^{m,n}$

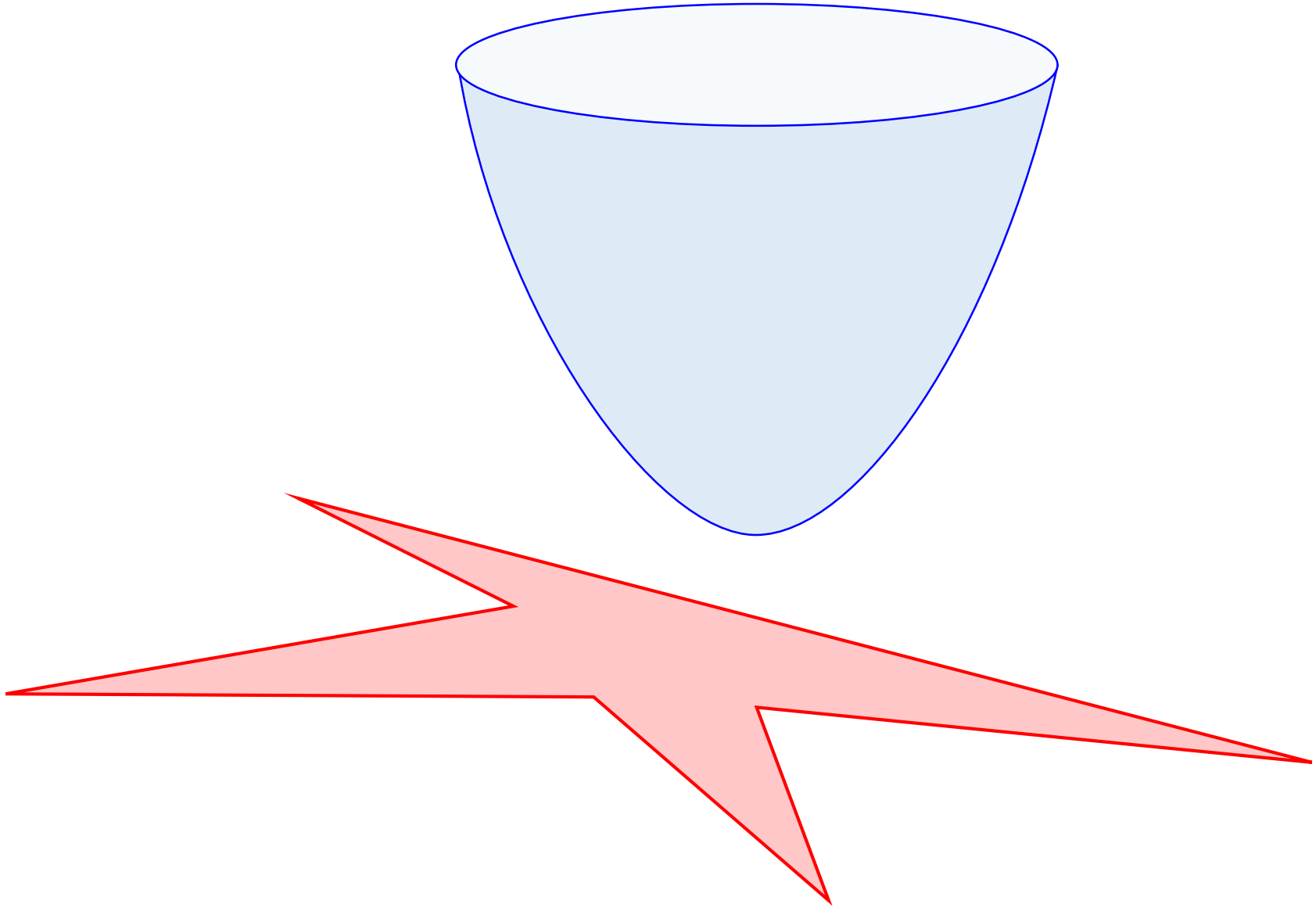


Perform k-truncated SVD

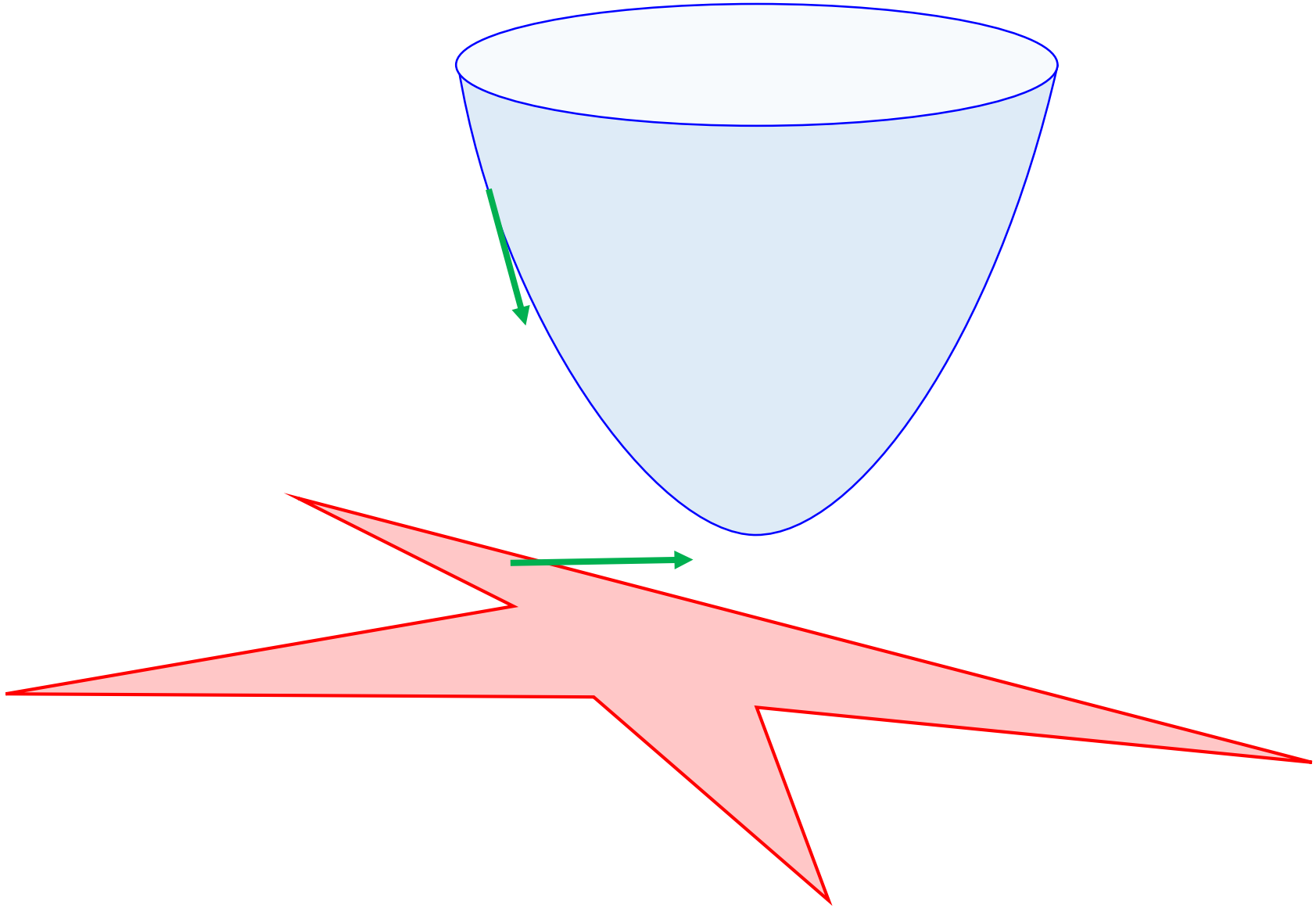
Sparse Recovery

$$\min_{\mathbf{w} \in \mathcal{B}_0^p(s)} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2$$

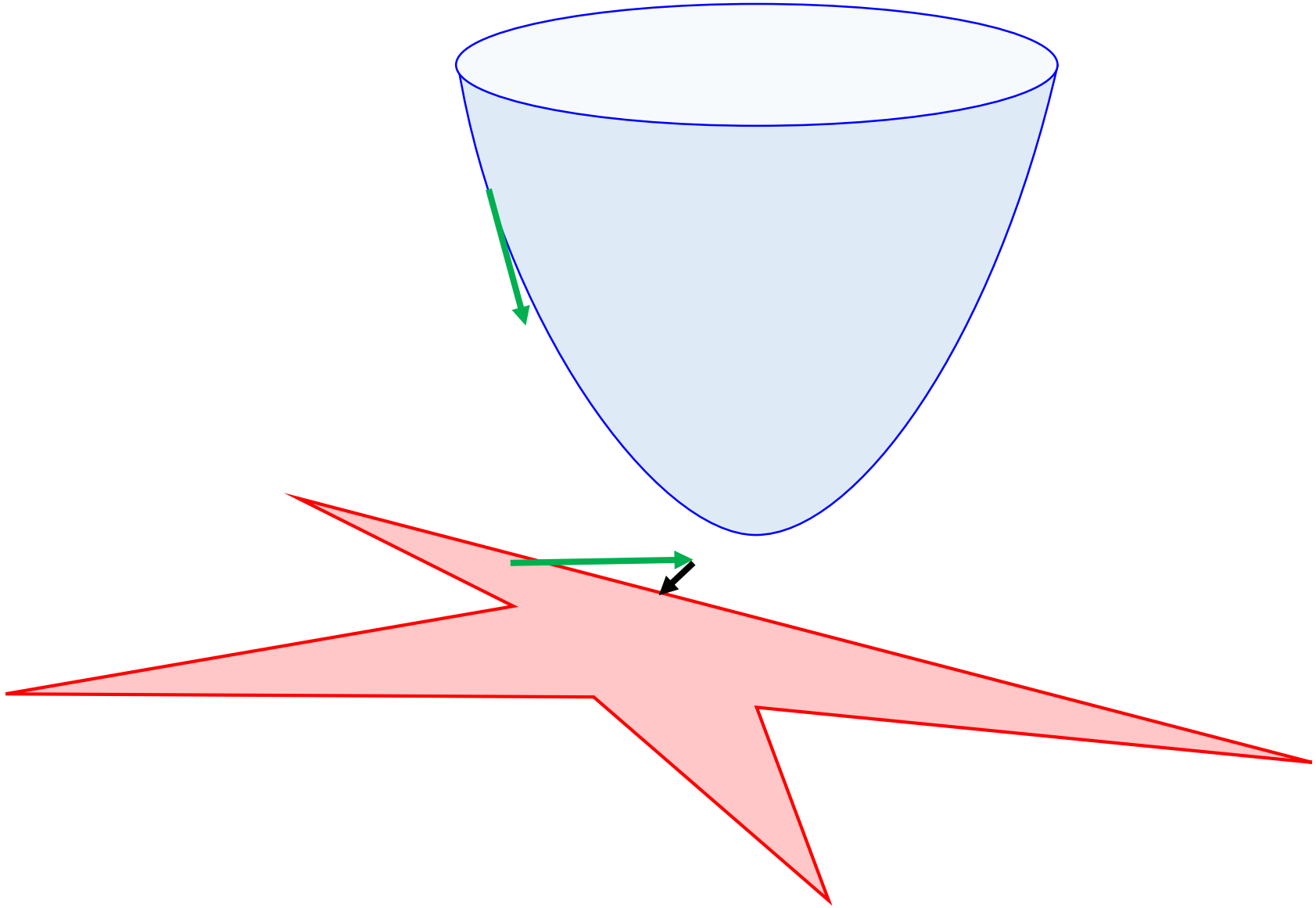
Projected Gradient Descent



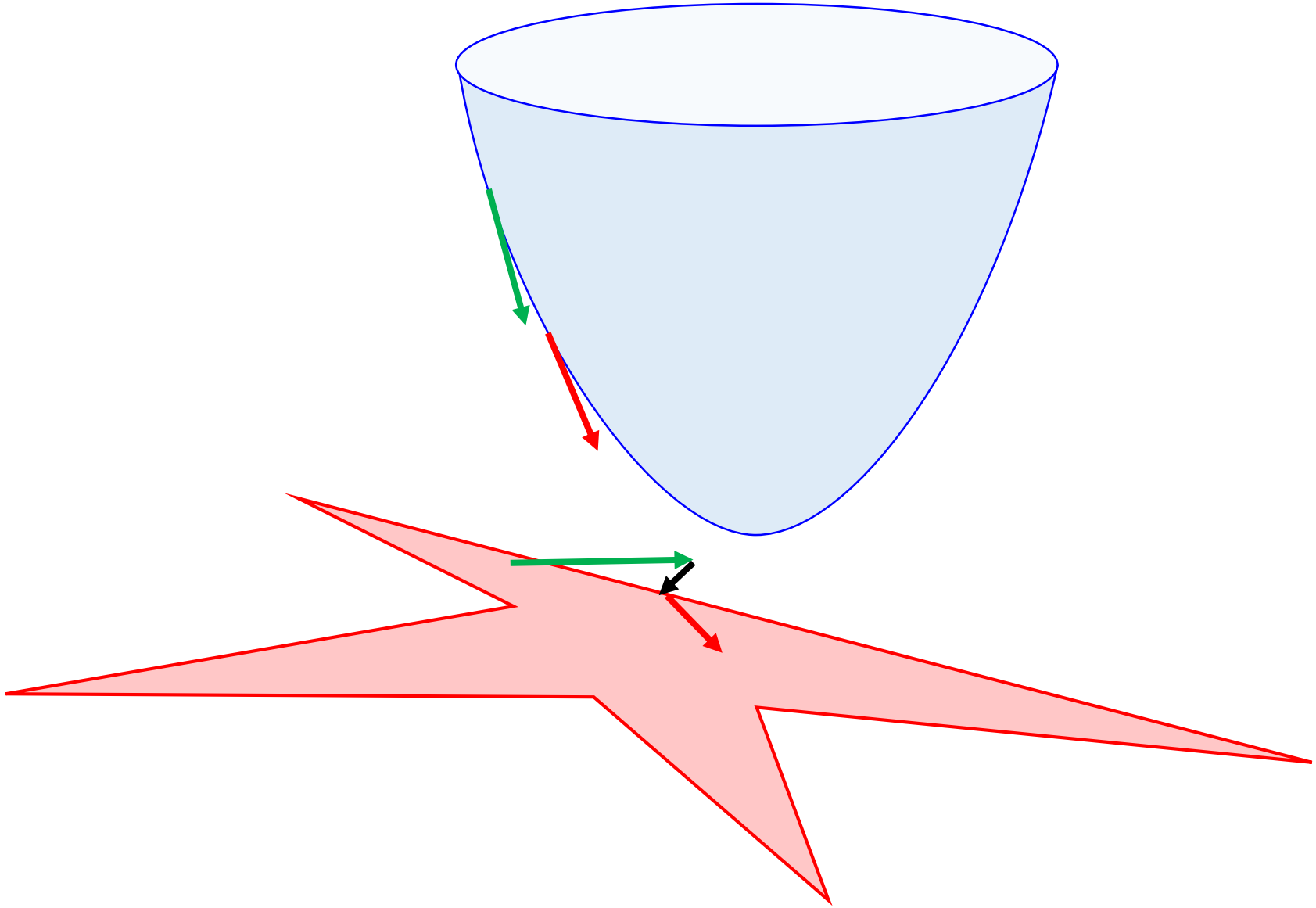
Projected Gradient Descent



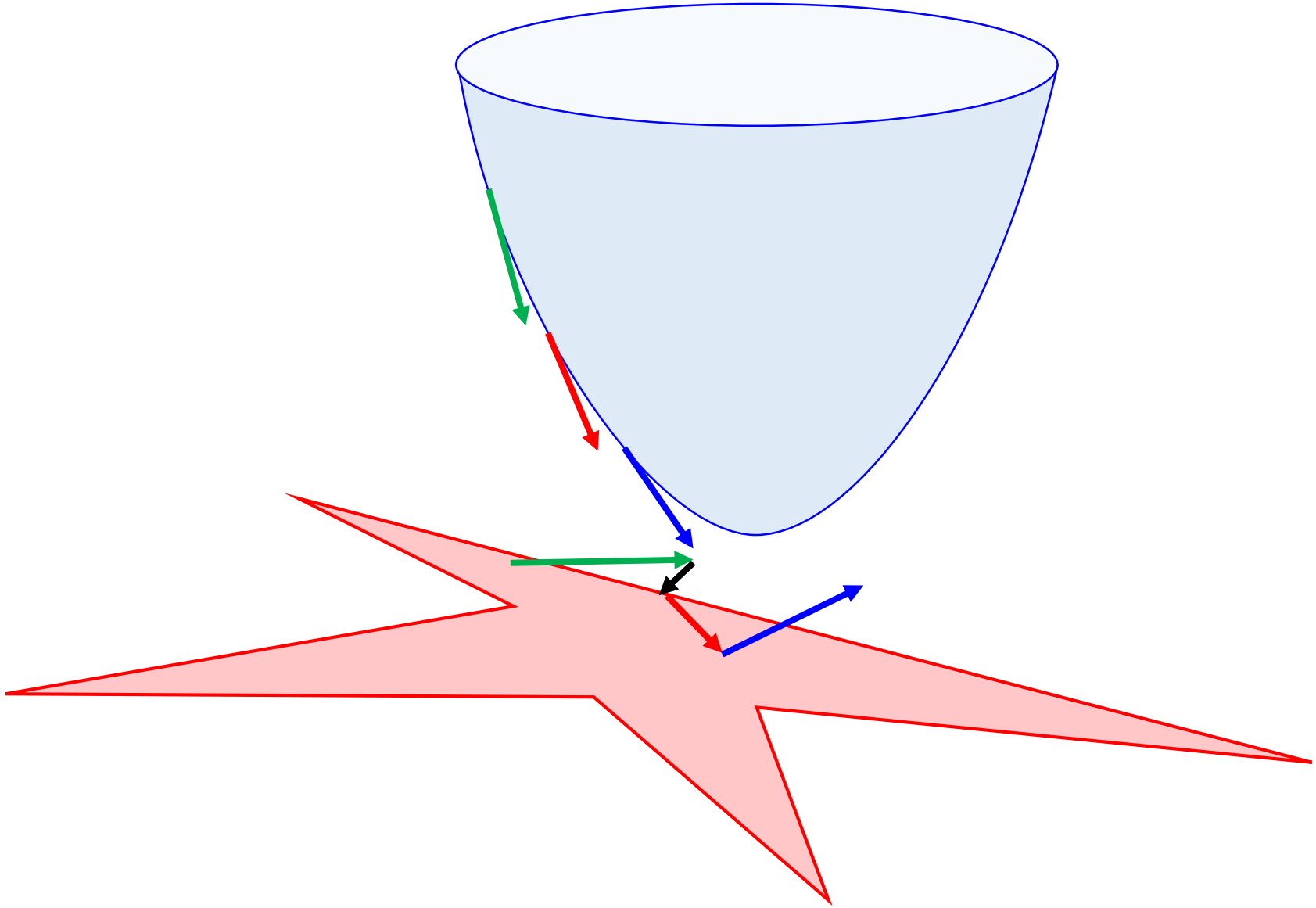
Projected Gradient Descent



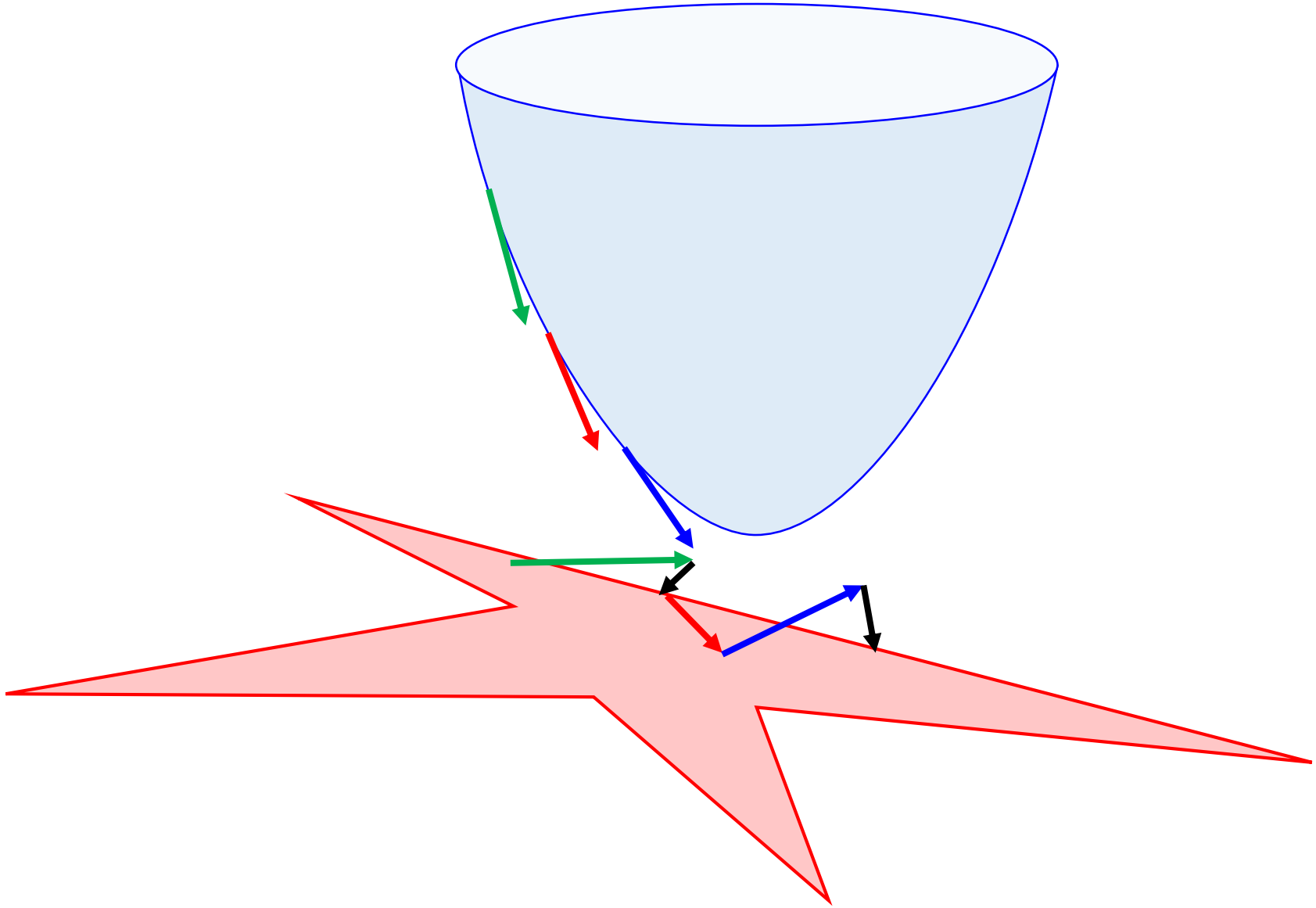
Projected Gradient Descent



Projected Gradient Descent



Projected Gradient Descent

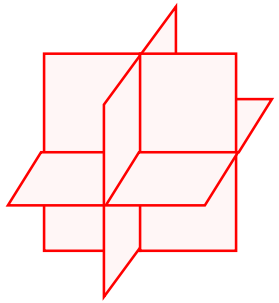


Pursuit and Greedy Methods

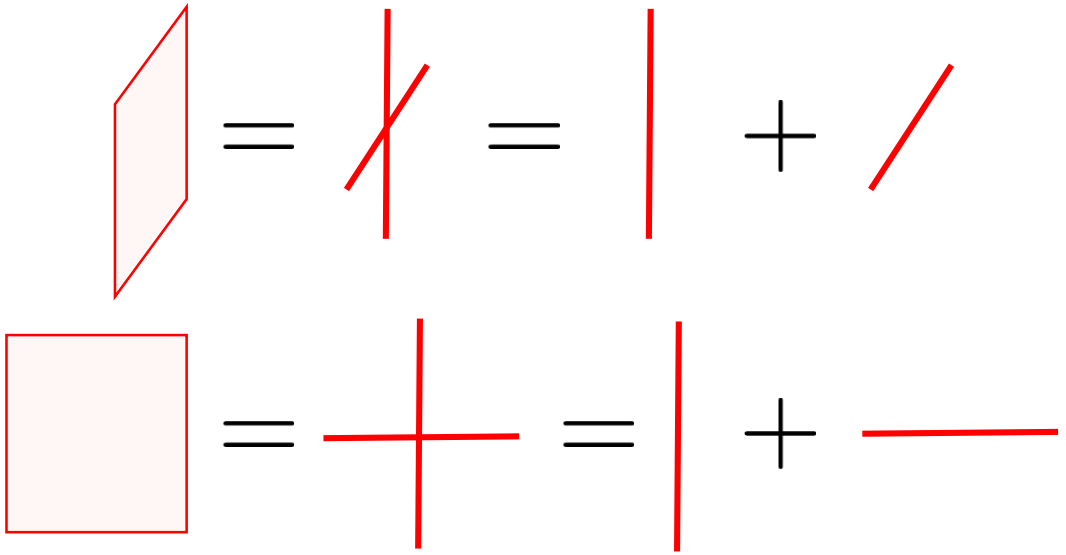
$$\min \ f(\mathbf{x})$$
$$s.t. \ \mathbf{x} \in \mathcal{C}$$

$\mathcal{A}$ Set of "atoms"

$$\mathcal{C} = \left\{ \mathbf{x} = \sum_{i=1}^{s} \mathbf{a}_i : \mathbf{a}_i \in \mathcal{A} \right\}$$



Sparse Recovery



- ▷ Initialize $S^0 = \phi$
- ▷ For $t = 1, 2, \ldots$
  - ▷ $\mathbf{a}^t =$ "best" greedy choice
  - ▷ $S^t = S^{t-1} \cup \{\mathbf{a}^t\}$
  - ▷ $\mathbf{x}^t = \underset{\mathbf{x} \in \mathrm{conv}(S^t)}{\arg\min} \ f(\mathbf{x})$

Here comes the pudding!

Foreground-background Separation

Convex Relaxation. Runtime: 1700 sec

AltMin. Runtime: 10 sec

[Netrapalli *et al* 2014]

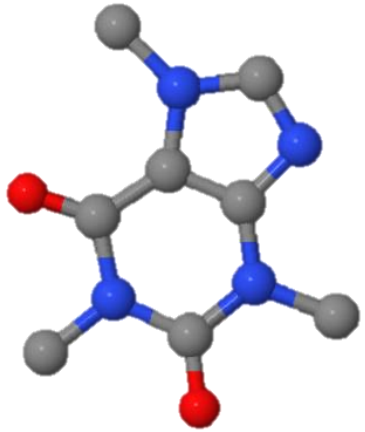
Image Retrieval from Spectral Measurements (CDF 20x3)

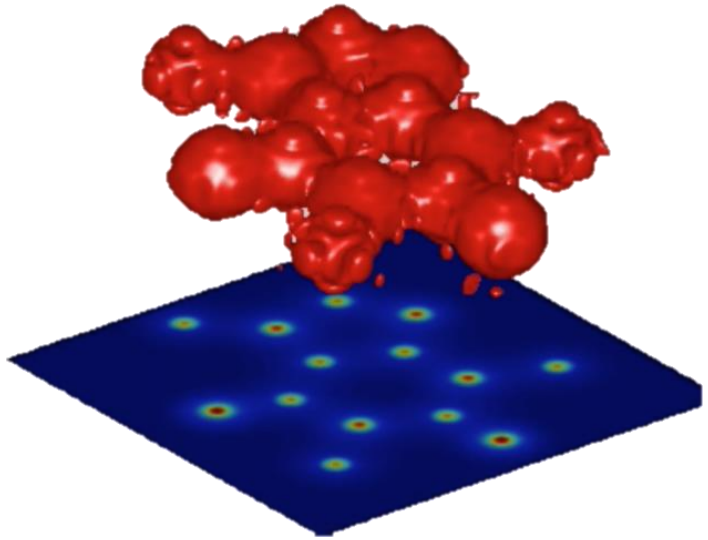
Naqsh-e Jahan Square, Esfahan (189 × 768 pixels), 60 sec, relative error 1e-16, SDP relaxation: 85GB

Milky way Galaxy (1080 × 1920 pixels), 21 min, relative error 1e-16, SDP relaxation: 17TB

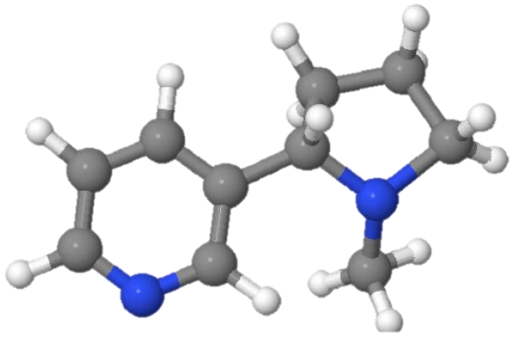
[Candes *et al* 2015]

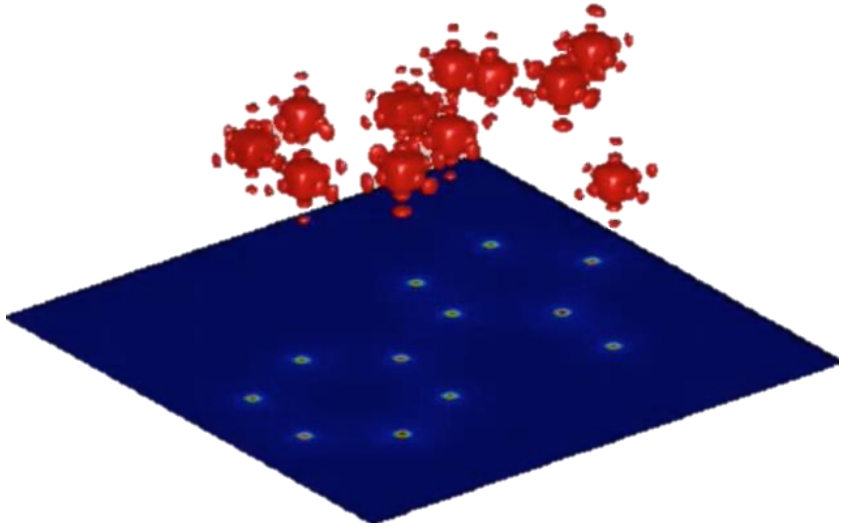
X-ray Crystallography (CDF 20)





Caffeine molecule, relative error 1e-6, 5.4 hours

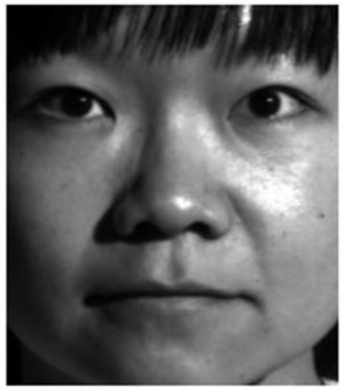




Nicotine molecule, relative error 1e-5, 5.4 hours

[Candes *et al* 2015]

Image Reconstruction

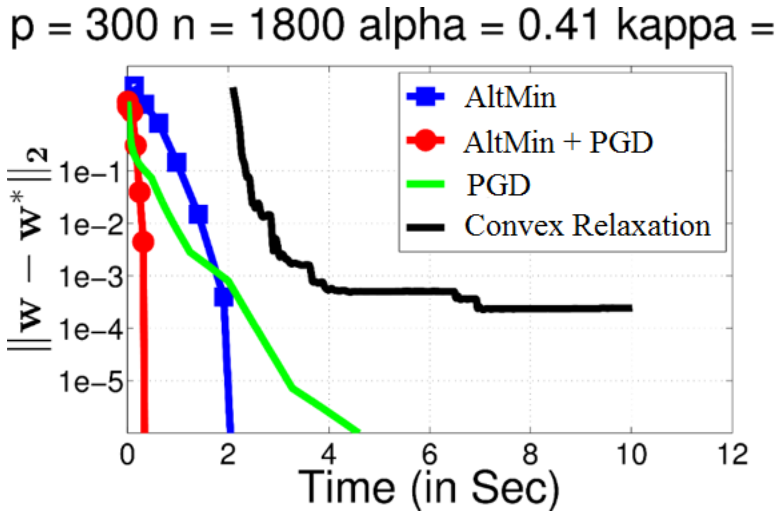


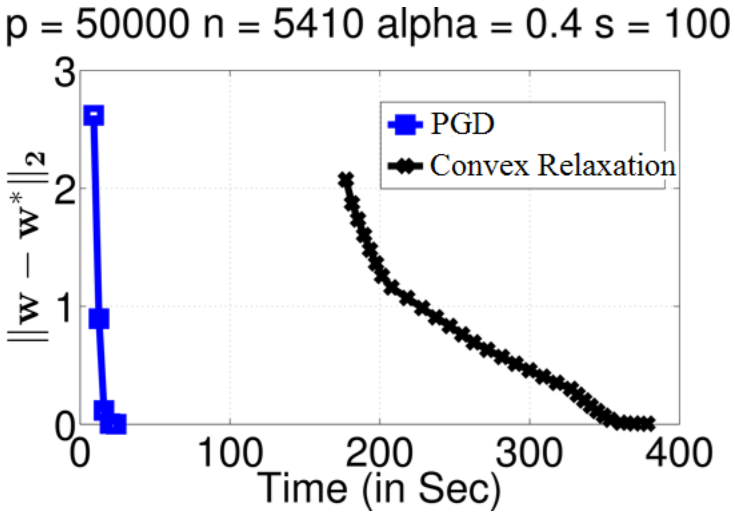
Original

Input

OLS

AltMin+PGD

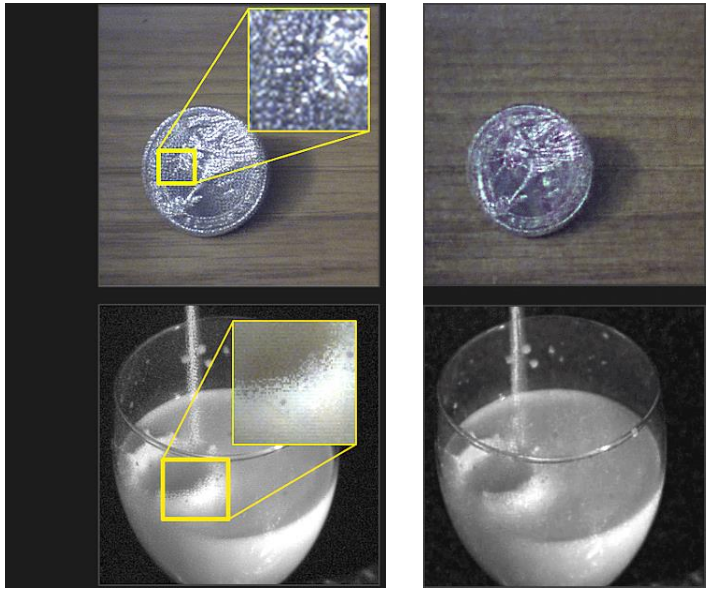


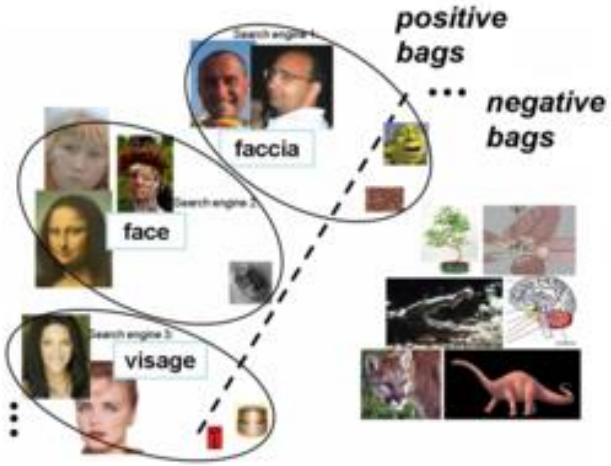


[Bhatia *et al* 2015]

Other Applications

- Image Superresolution
- Image Inpainting
- High-speed videography
- Multi-instance Learning
- Latent Variable Models





safaribooksonline.com, newatlas.com, cs.texas.edu

Questions?