Tight Bounds for the Approximate Carathéodory Theorem

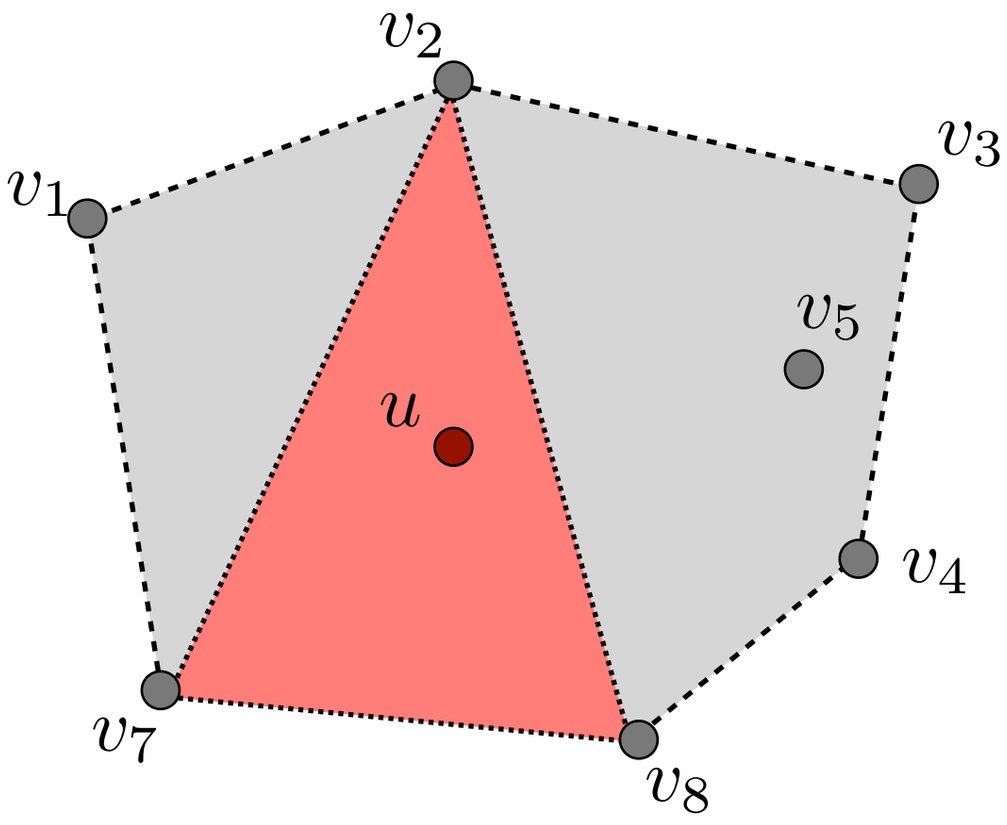
Vahab Mirrokni
Google

Renato Paes Leme
Google

Adrian Vladu
MIT

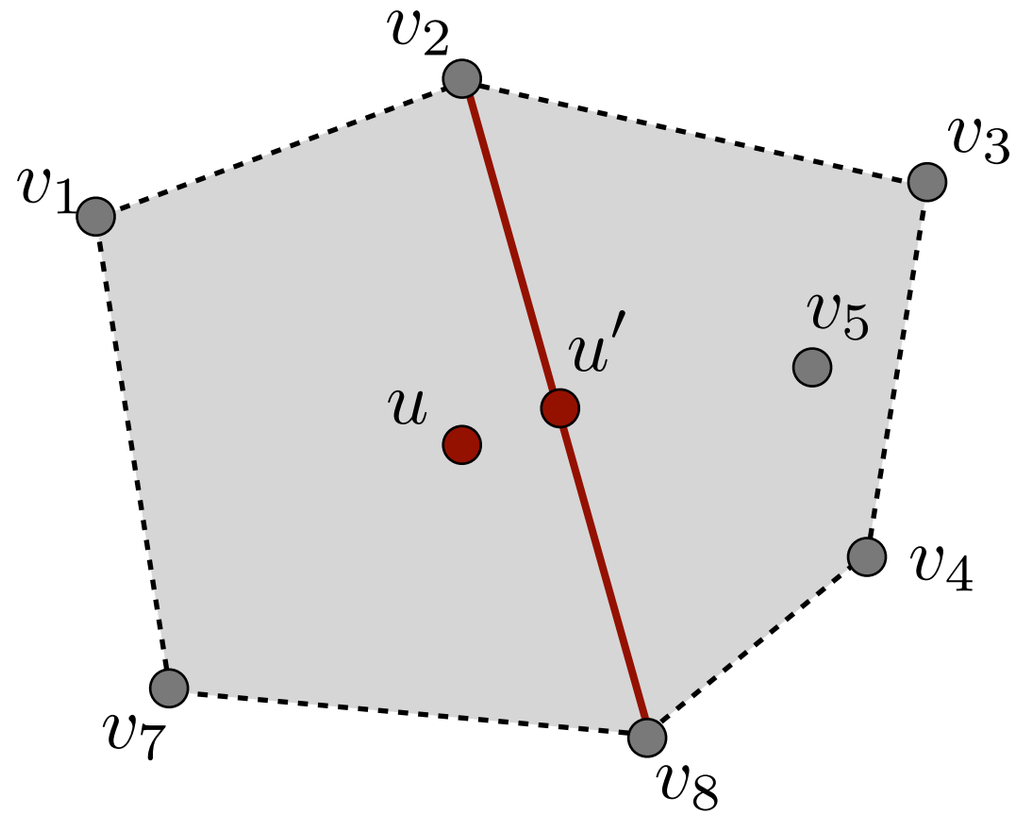
Sam Chiu-wai Wong
UC Berkeley

Exact Carathéodory Theorem



Given a collection of points $V \subseteq \mathbb{R}^d$ and a point u in the convex hull, then u is in the convex combination of d+1 points of V .

Approximate Carathéodory Theorem



Given a collection of points $V \subseteq \mathbb{R}^d$ and a point u in the convex hull, then for $p \geq 2$ and $v_i, u \in B_p(1)$ there is u' in the convex hull of $4p/\epsilon^2$ points of V with $\|u - u'\|_p \leq \epsilon$.

Brief History of Approx Caratheodory

- [Barman, STOC'15]: there exist $k = O(p/\epsilon^2)$ points such that $\frac{1}{k}\sum_{i=1}^{k} v_i \approx u$

  - *Application*: ϵ-nets for $V\Delta = \{Vx; x \in \Delta\}, V = [v_1, \ldots, v_n]$
    There is a set of $n^{O(p/\epsilon^2)}$ that approximate $V\Delta$ in the ℓ_p-norm.

  - *Bilinear programs*: programs of the type $\max y^\top Vx + a^\top x + b^\top y$ s.t. ...
    can be solved by enumerating over possible values of Vx .

- PTAS for Nash equilibrium in s-sparse bi-matrix games $n^{O(\log s/\epsilon^2)}$

- additive PTAS for the k-densest subgraph problem for bounded degree.

- applications in combinatorics

- lower bound of $k \geq \Omega(1/\epsilon^{p/(p-1)})$. So the result is tight for ℓ_2 .

Brief History of Approx Carathéodory

- [Barman, STOC'15]: there exist $k = O(p/\epsilon^2)$ points such that $\frac{1}{k}\sum_{i=1}^{k} v_i \approx u$
- [Maurey, 1980]: functional analysis
  - [Shalev-Shwartz, Srebro and Zhang, 2010]:
    sparsity / accuracy tradeoffs in linear regression
- [Novikoff, 1962]: analysis of the perceptron algorithm implies an ℓ_2 -version.

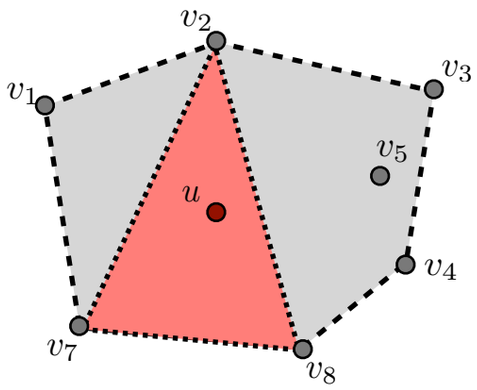
Brief History of Approx Carathéodory

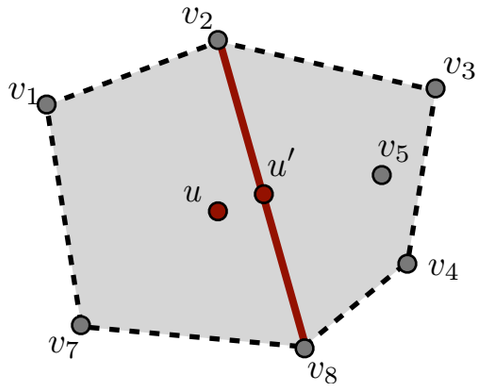
- [Barman, STOC'15]: there exist $k = O(p/\epsilon^2)$ points such that $\frac{1}{k}\sum_{i=1}^{k} v_i \approx u$
- [Maurey, 1980]: functional analysis
  - [Shalev-Shwartz, Srebro and Zhang, 2010]: sparsity / accuracy tradeoffs in linear regression
- [Novikoff, 1962]: analysis of the perceptron algorithm implies an ℓ_2 -version.

This paper:

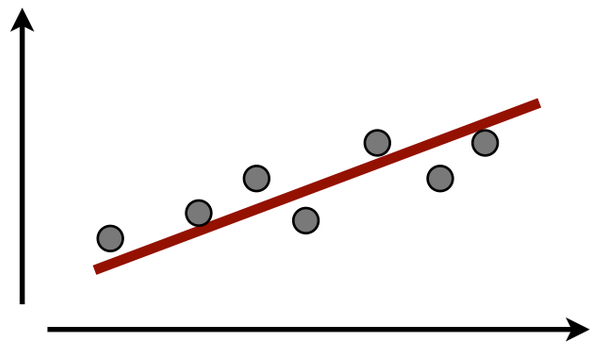
- A deterministic $O(Np/\epsilon^2)$ algorithms via optimization.
- A lower bound showing that $O(p/\epsilon^2)$ is tight.

Sparsification via Optimization



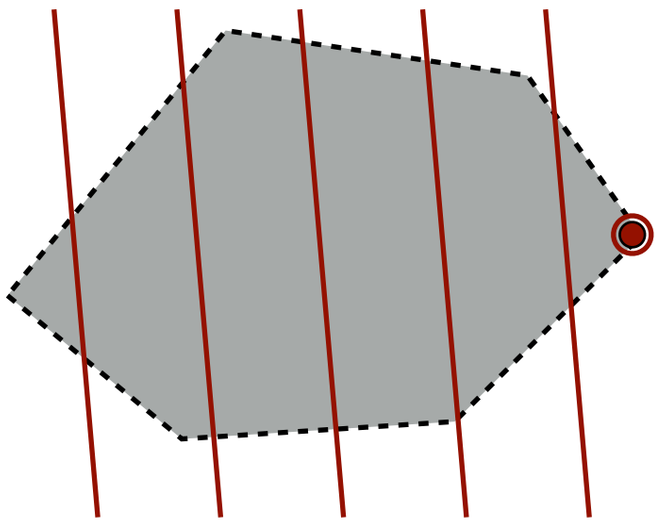


Convex combinations



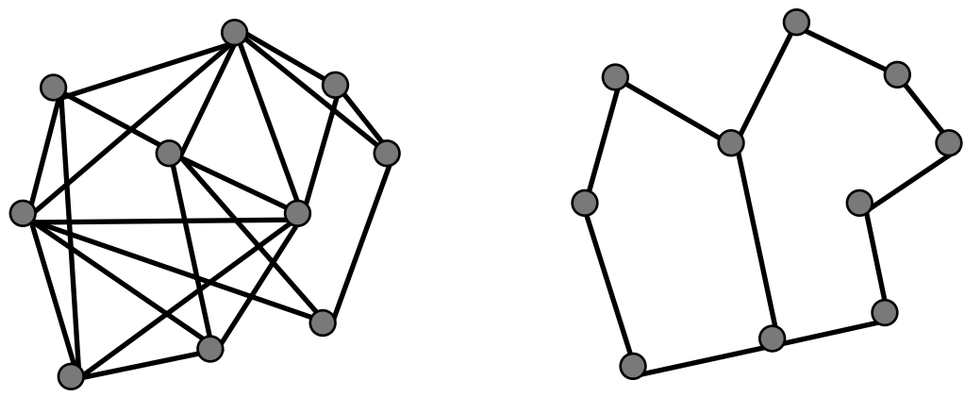
Linear regressions

$$\min \|Ax - b\|_2^2$$



Linear Programming

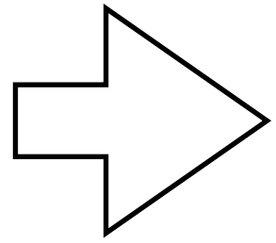
$$\max c^\top x \text{ s.t. } Ax \leq b$$



Graph sparsification

$$(1-\epsilon)G \preccurlyeq H \preccurlyeq (1+\epsilon)G$$

Sparsification via Optimization

Exact solution  Approximate optimal solution

Plan #1:

(1) solve the exact problem.

(2) sample from it.

Sparsification via sampling

- Barman's proof of the Approximate Caratheodory Theorem

- First solve the exact Caratheodory problem: $u = \sum_i x_i \cdot v_i$
  for $x \in \Delta := \left\{x \in \mathbb{R}^n_+; x_i \geq 0; \sum_i x_i = 1\right\}$

- Interpret x as a probability distribution over vectors $\{v_1, \ldots, v_n\}$

- Sample k vectors according to those probabilities.

- Use concentration bounds to argue that $\frac{1}{k}\sum_{i=1}^k \hat{v}_i$ is close to u.

  More precisely, use Khintchine's inequality to bound $\mathbb{E}\left\|u - \frac{1}{k}\sum_{i=1}^k \hat{v}_i\right\|_p \leq \epsilon$

- for $p \to \infty$ use that $\|x\|_\infty \leq \|x\|_{\log n} \leq n^{1/\log n} \cdot \|x\|_\infty = O(1) \cdot \|x\|_\infty$
  and apply Chernoff bounds + union bound.

Sparsification via sampling (digression)

- **Games** [Lipton, Young], [Lipton, Markakis, Mehta]: There are ϵ -Nash equilibrium where each player employs a mixed strategy with support $O(\log n/\epsilon^2)$

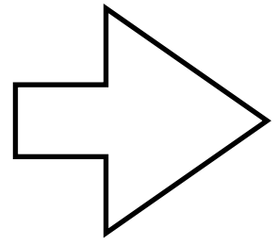
- **Linear Programs**: Positive linear programs $\max c^\top x \text{ s.t. } Ax \leq b$ can be converted to $\min_{x\in\Delta} \|\tilde{A}x\|_\infty$ which is a zero sum game $\min_{x\in\Delta}\max_{y\in\Delta} y^\top \tilde{A}x$, so there exist sparse approximate primal-dual pairs.

- **Graph sparsification** [Spielman, Srivastava]: The Laplacian can be written as the sum of Laplacians of the edges $L_G = \sum_e L_e$. We would like to sample edges so that $L_G \approx \sum_{i=1}^k L_{e_k}$ in the spectral norm. After a suitable normalization:

$$I = \sum_e \tau_e \left(\tfrac{1}{\tau_e} L_G^{+/2} L_e L_G^{+/2} \right)$$

  we can sample according to effective resistances τ_e and apply matrix Chernoff bounds.

Sparsification via Optimization

Exact solution  Approximate optimal solution

Plan #1:

(1) solve the exact problem.

(2) sample from it.

(1) deterministic
(2) by-passes the exact problem

Plan #2:

(1) write as a convex optimization problem.

(2) pass to the dual.

(3) k steps of gradient / mirror descent.

Carathéodory from optimization

- There is a natural convex function to minimize:

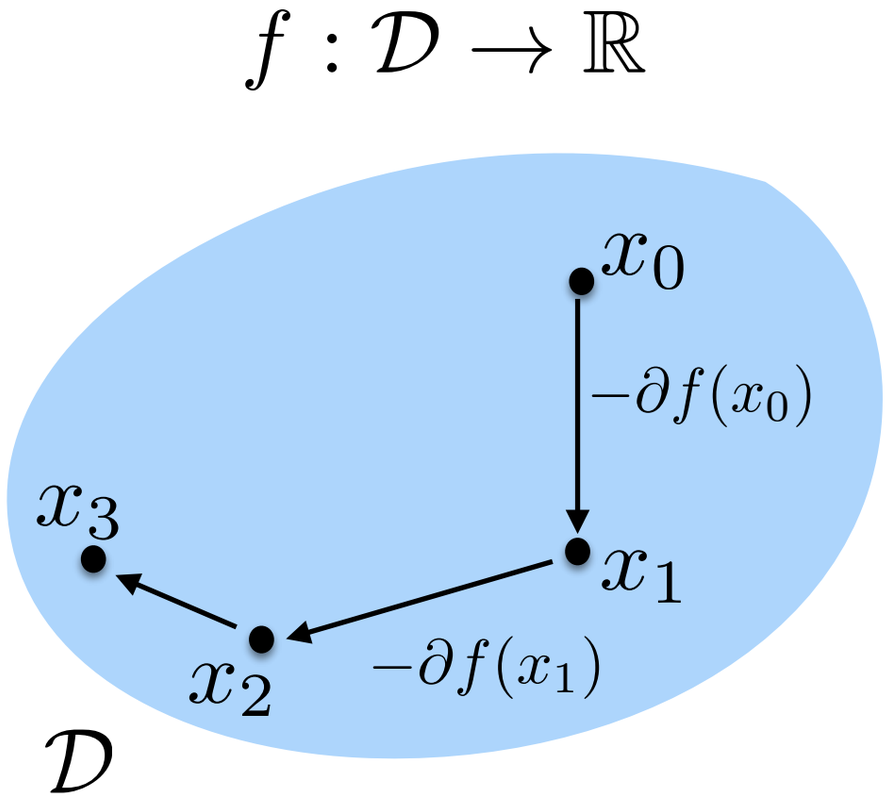
$$\min_{x \in \Delta} \|Vx - u\|_p$$

Carathéodory from optimization

- There is a natural convex function to minimize:

$$\min_{x \in \Delta} \underbrace{\|Vx - u\|_p}_{f(x)}$$

- A natural idea is to follow gradients.



Gradient Descent:

- Update: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$
- Guarantee: $f(x_t) - f(x^*) \leq \frac{RL}{\sqrt{t}}$
- R = diameter of the domain
- L = Lipschitz constant

Carathéodory from optimization

- There is a natural convex function to minimize:

$$\min_{x \in \Delta} \underbrace{\|Vx - u\|_p}_{f(x)}$$

- A natural idea is to follow gradients.

Problems with this approach:

- Gradient is not sparse:

$$\nabla f(x) = \frac{V^\top V x - V^\top u}{\|Vx - u\|_2}$$

- No good bound to $\|\nabla f(x)\|_2$

Gradient Descent:

- Update: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$
- Guarantee: $f(x_t) - f(x^*) \leq \frac{RL}{\sqrt{t}}$
- R = diameter of the domain
- L = Lipschitz constant

Carathéodory from optimization

- The first idea is to pass to the dual:

  Primal formulation: $\min_{x \in \Delta} \|Vx - u\|_p$

Carathéodory from optimization

- The first idea is to pass to the dual:

  Primal formulation: $\min_{x \in \Delta} \|Vx - u\|_p$

  Saddle point formulation: $\min_{x \in \Delta} \max_{y \in B_q(1)} y^\top (Vx - u)$

$$B_q(1) := \{y \in \mathbb{R}^n; \|y\|_q \leq 1\} \quad \text{for} \quad \tfrac{1}{p} + \tfrac{1}{q} = 1$$

$$\|x\|_p = \max_{y \in B_q(1)} y^\top x$$

Carathéodory from optimization

- The first idea is to pass to the dual:

  Primal formulation: $\min_{x \in \Delta} \|Vx - u\|_p$

  Saddle point formulation: $\min_{x \in \Delta} \max_{y \in B_q(1)} y^\top (Vx - u)$

  Dual formulation: $\max_{y \in B_q(1)} g(y) := \min_{x \in \Delta} y^\top (Vx - u)$

  Sion's Theorem

Carathéodory from optimization

- The first idea is to pass to the dual:

  Primal formulation: $\min_{x \in \Delta} \|Vx - u\|_p$

  Saddle point formulation: $\min_{x \in \Delta} \max_{y \in B_q(1)} y^\top (Vx - u)$

  Dual formulation: $\max_{y \in B_q(1)} g(y) := \min_{x \in \Delta} y^\top (Vx - u)$

Sion's Theorem

Sion's Theorem: Given X convex and compact, Y convex and a function f(x,y) convex in x and concave in y, then:

$$\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y)$$

Von Neumann: $X = Y = \Delta$ and f bilinear.

Carathéodory from optimization

- The first idea is to pass to the dual:

  Primal formulation: $\min_{x \in \Delta} \|Vx - u\|_p$

  Saddle point formulation: $\min_{x \in \Delta} \max_{y \in B_q(1)} y^\top (Vx - u)$

  Dual formulation: $\max_{y \in B_q(1)} g(y) := \min_{x \in \Delta} y^\top (Vx - u)$

- Gradients are now very nice:

Envelope Theorem: if $f(x) = \min_i f_i(x)$ then $\nabla f(x) = \nabla f_i(x)$ for $i \in \operatorname{argmin}_i f_i(x)$.

Carathéodory from optimization

- The first idea is to pass to the dual:

  Primal formulation: $\min_{x \in \Delta} \|Vx - u\|_p$

  Saddle point formulation: $\min_{x \in \Delta} \max_{y \in B_q(1)} y^\top (Vx - u)$

  Dual formulation: $\max_{y \in B_q(1)} g(y) := \min_{x \in \Delta} y^\top (Vx - u)$

- Gradients are now very nice: $g(y) = y^\top (Ve_i - u)$ so: $\nabla g(y) = v_i - u$

Envelope Theorem: if $f(x) = \min_i f_i(x)$ then $\nabla f(x) = \nabla f_i(x)$ for $i \in \operatorname{argmin}_i f_i(x)$.

Carathéodory from optimization

- The first idea is to pass to the dual:

  Primal formulation: $\min_{x \in \Delta} \|Vx - u\|_p$

  Saddle point formulation: $\min_{x \in \Delta} \max_{y \in B_q(1)} y^\top (Vx - u)$

  Dual formulation: $\max_{y \in B_q(1)} g(y) := \min_{x \in \Delta} y^\top (Vx - u)$

- Gradients are now very nice: $g(y) = y^\top (Ve_i - u)$ so: $\nabla g(y) = v_i - u$

- Gradient descent guarantees that

$$\left\| \tfrac{1}{T} \textstyle\sum_t \nabla g(x_t) \right\|_p \leq \tfrac{1}{T} \textstyle\sum_t \nabla g(x_t)^\top y_t + \tfrac{RL}{\sqrt{T}} \leq \tfrac{RL}{\sqrt{T}}$$

Carathéodory from optimization

- The first idea is to pass to the dual:

  Primal formulation: $\min_{x \in \Delta} \|Vx - u\|_p$

  Saddle point formulation: $\min_{x \in \Delta} \max_{y \in B_q(1)} y^\top (Vx - u)$

  Dual formulation: $\max_{y \in B_q(1)} g(y) := \min_{x \in \Delta} y^\top (Vx - u)$

- Gradients are now very nice: $g(y) = y^\top (Ve_i - u)$ so: $\nabla g(y) = v_i - u$

- Gradient descent guarantees that

$$\left\| \tfrac{1}{T} \textstyle\sum_t \nabla g(x_t) \right\|_p \leq \tfrac{1}{T} \textstyle\sum_t \nabla g(x_t)^\top y_t + \tfrac{RL}{\sqrt{T}} \leq \tfrac{RL}{\sqrt{T}}$$

$$=$$

$$\left\| \tfrac{1}{T} \textstyle\sum_t v_t - u \right\|_p$$

Almost there ! But $L = \|v_i - u\|_2 = O(n^{\frac{1}{2} - \frac{1}{p}})$

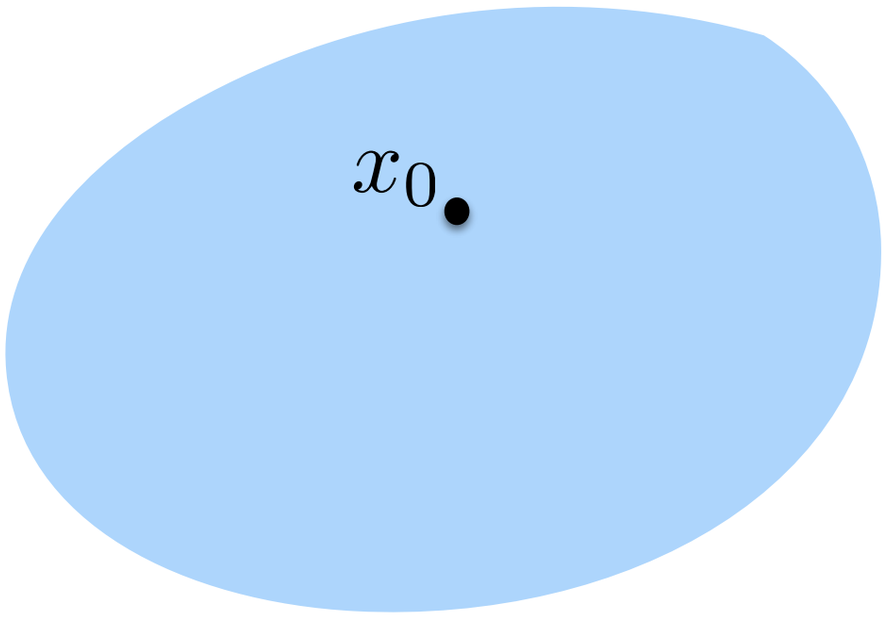
Mirror Descent

- Gradient Descent iteration: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$

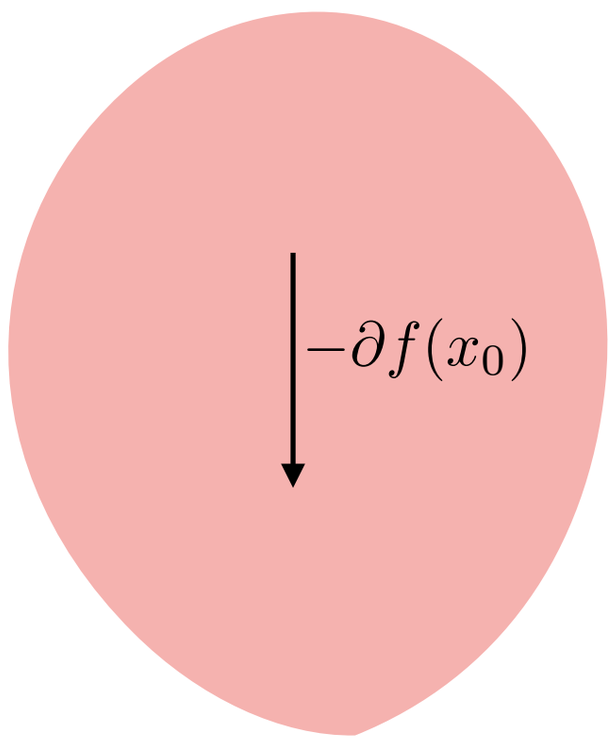
Mirror Descent

- Gradient Descent iteration: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$
- x_t : primal points, column vectors
- $\nabla f(x_t)$: linear forms (dual), row vectors

$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle$$



primal space

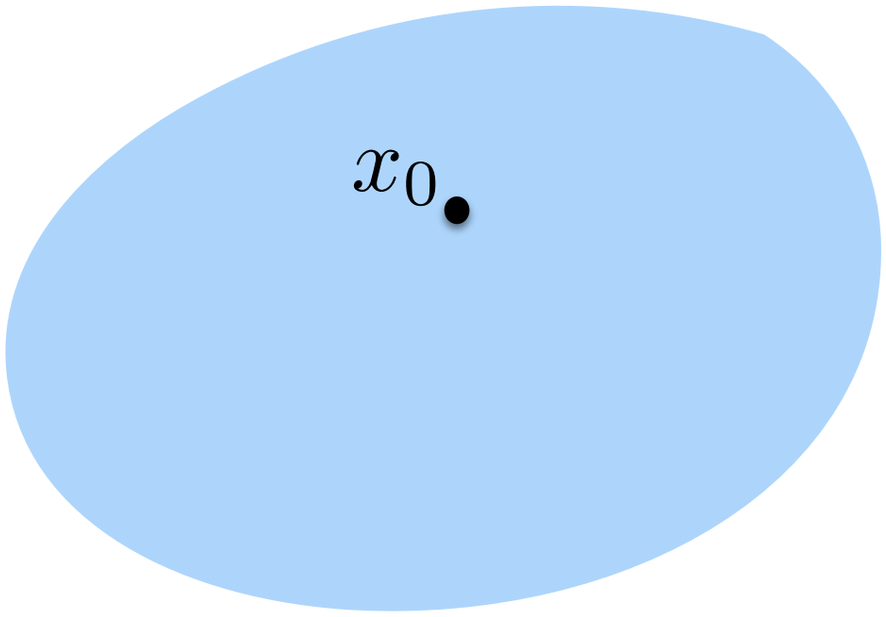


dual space

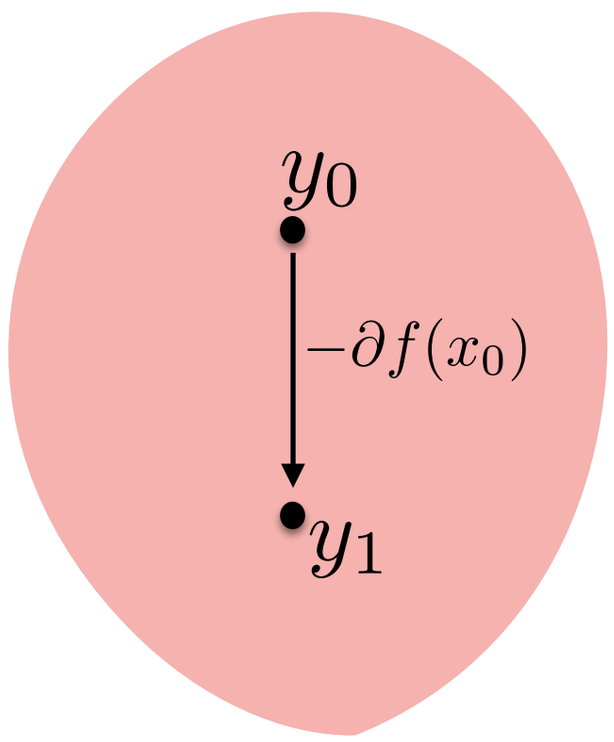
Mirror Descent

- Gradient Descent iteration: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$
- x_t : primal points, column vectors
- $\nabla f(x_t)$: linear forms (dual), row vectors

$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle$$



primal space



dual space

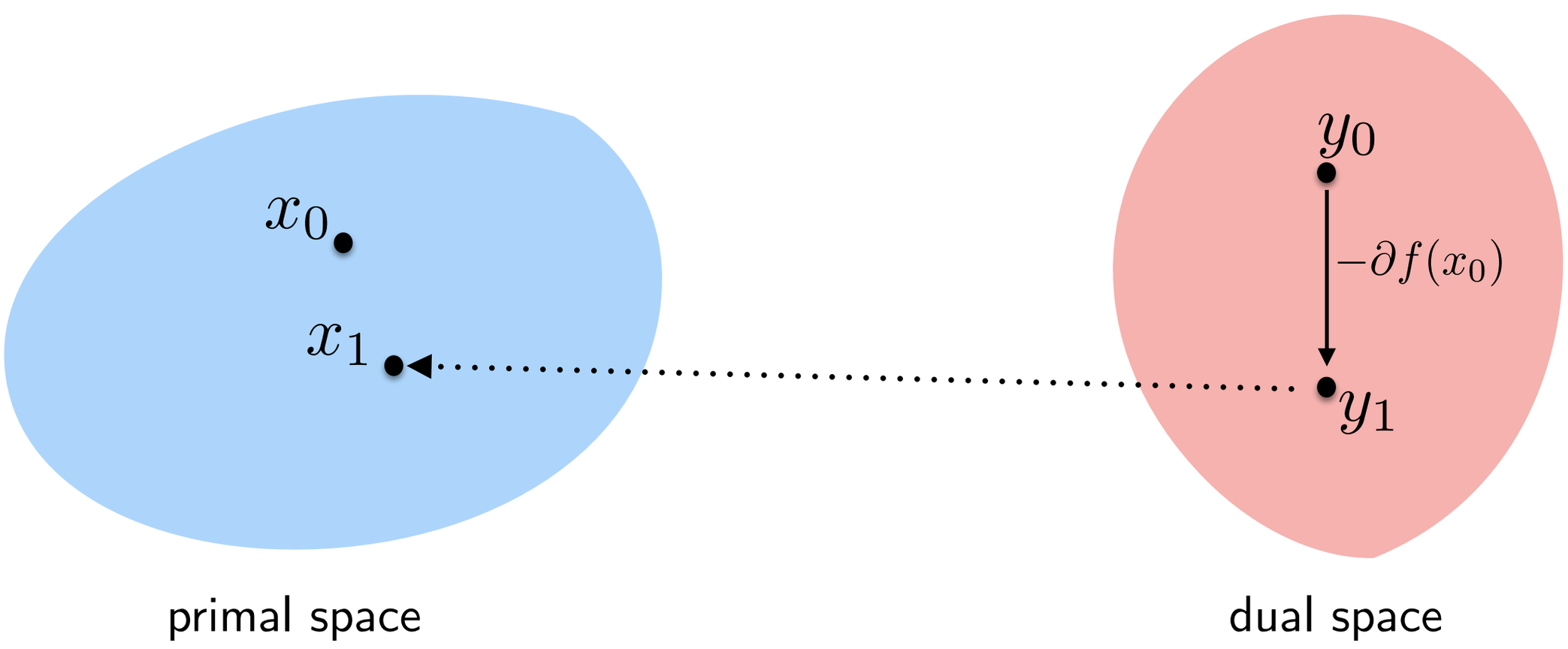
Mirror Descent

- Gradient Descent iteration: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$

- x_t : primal points, column vectors

$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle$$

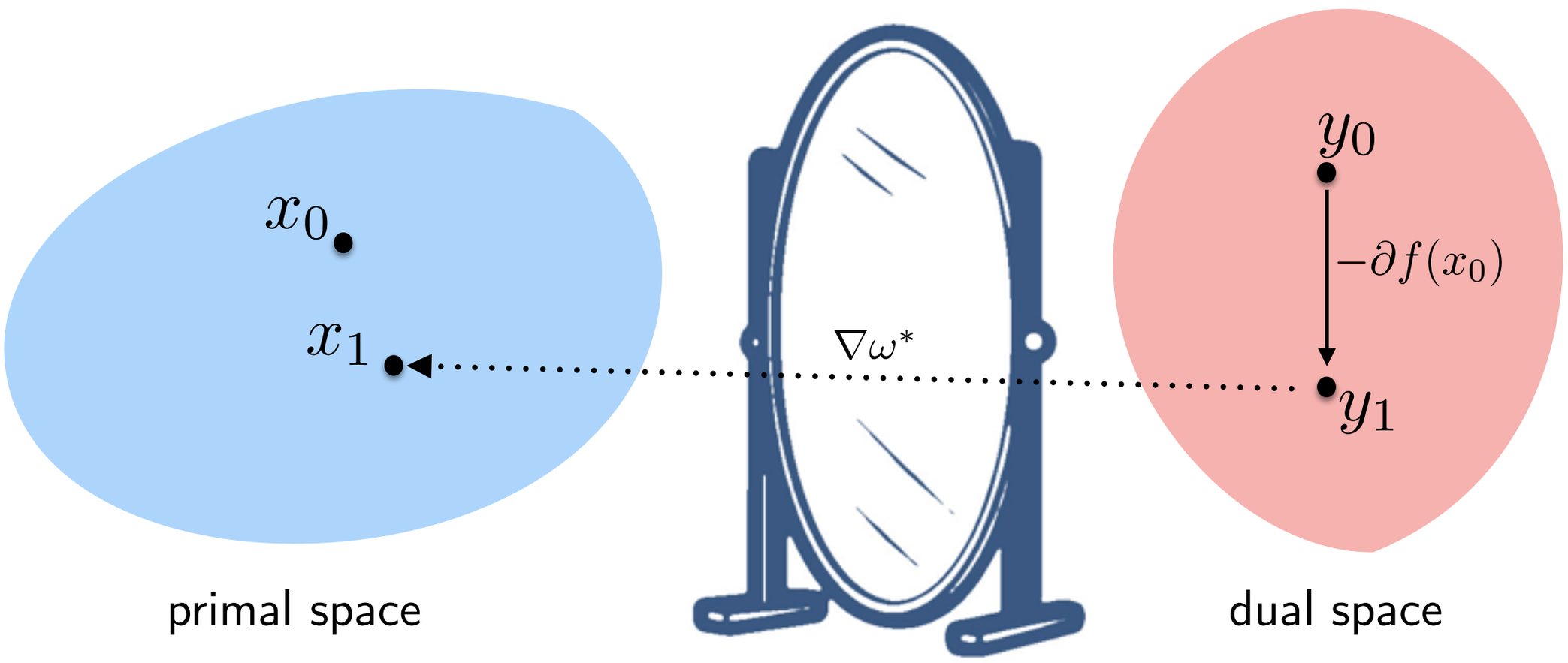
- $\nabla f(x_t)$: linear forms (dual), row vectors



Mirror Descent

- Gradient Descent iteration: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$
- x_t : primal points, column vectors
- $\nabla f(x_t)$: linear forms (dual), row vectors

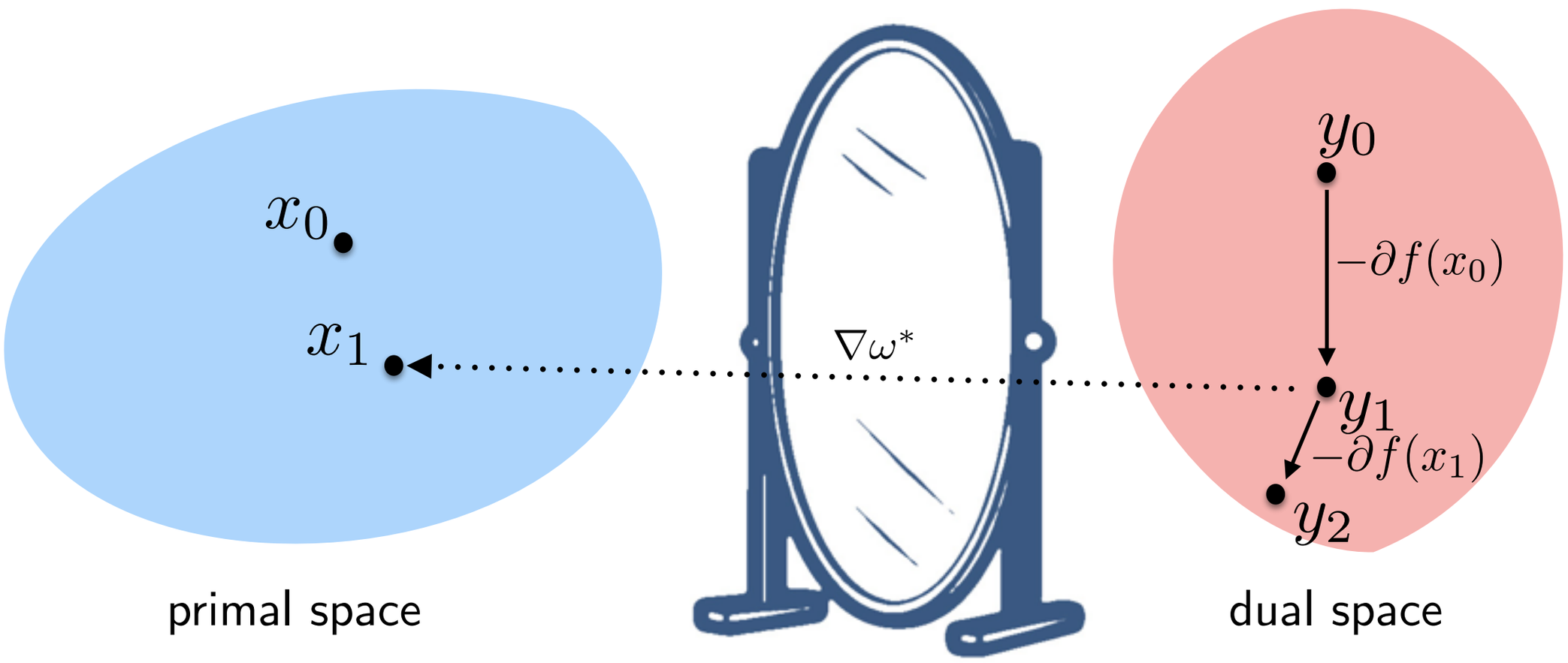
$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle$$



Mirror Descent

- Gradient Descent iteration: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$
- x_t : primal points, column vectors
- $\nabla f(x_t)$: linear forms (dual), row vectors

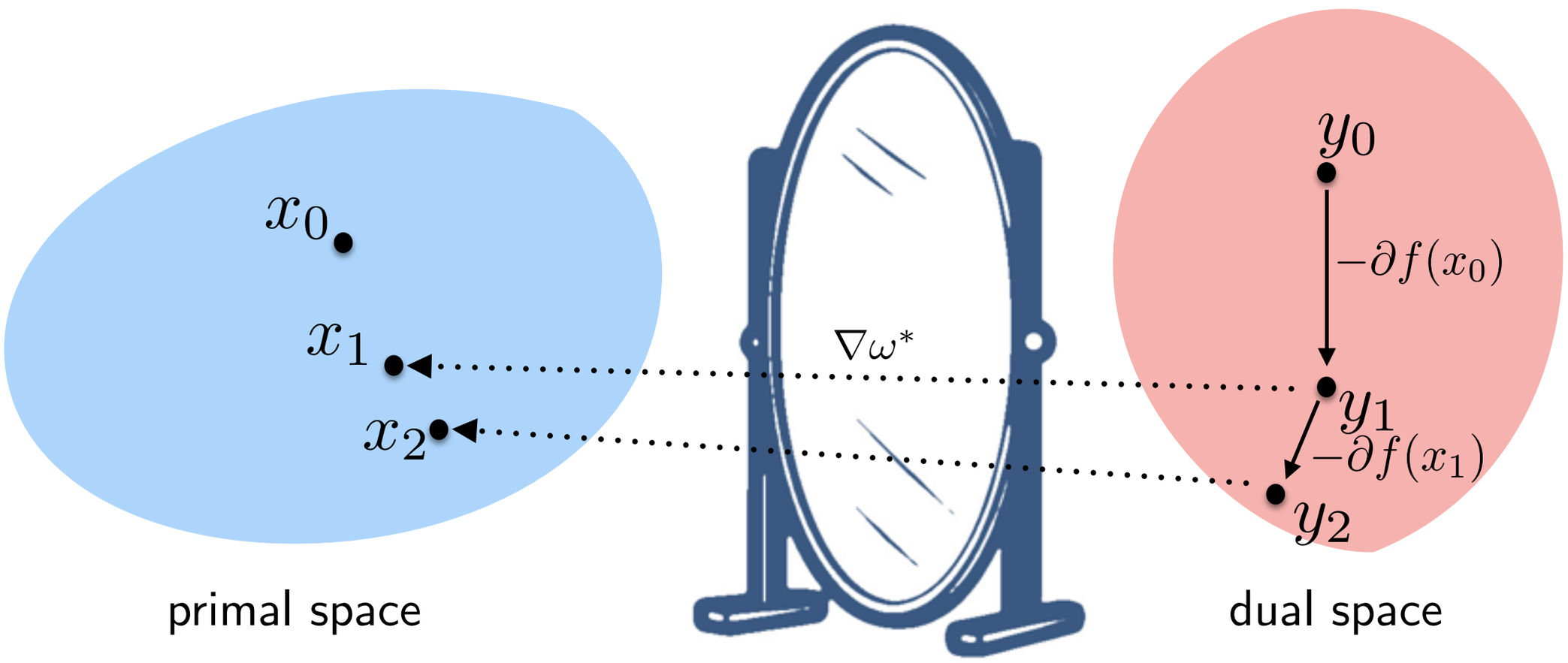
$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle$$



Mirror Descent

- Gradient Descent iteration: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$
- x_t : primal points, column vectors
- $\nabla f(x_t)$: linear forms (dual), row vectors

$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle$$

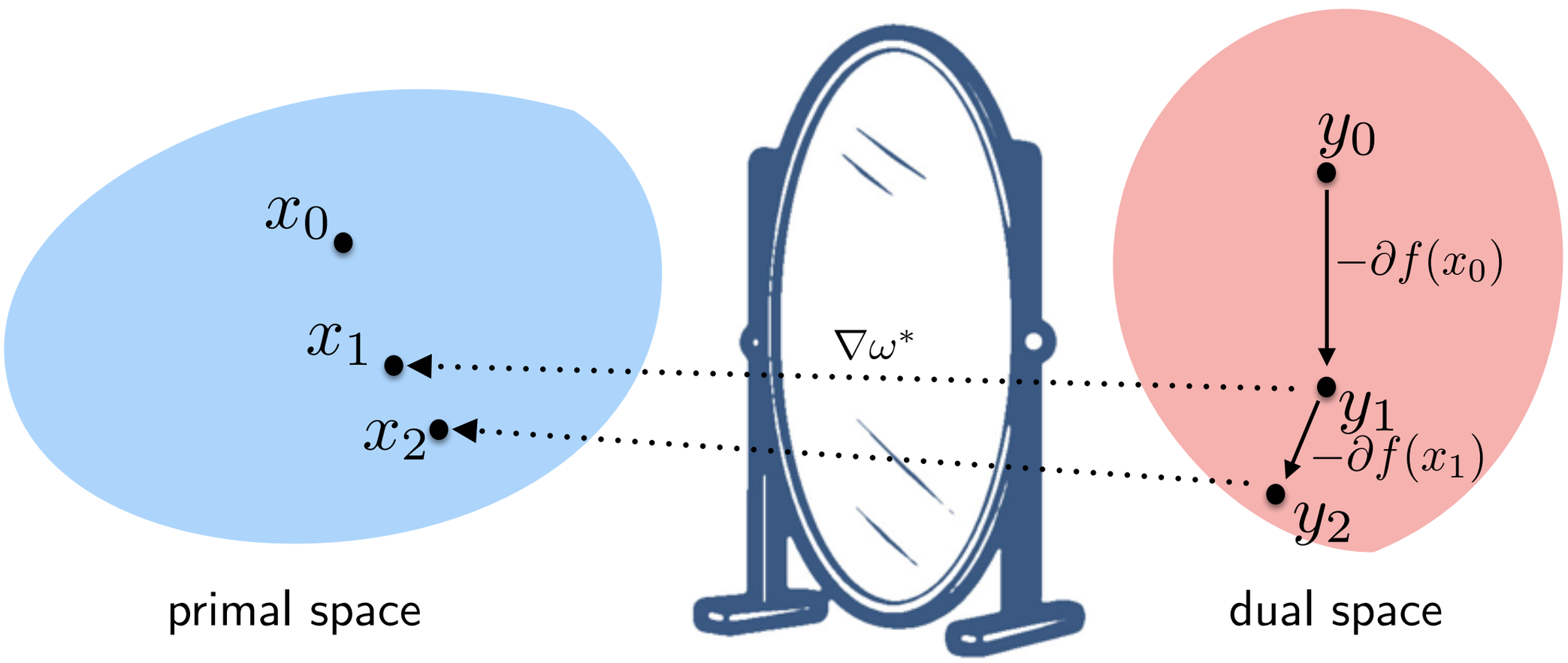


Mirror Descent

- Mirror Descent Iteration:

$$y_{t+1} = y_t - \eta_t \nabla f(x_t)$$

$$x_{t+1} = \nabla \omega^*(y_t)$$

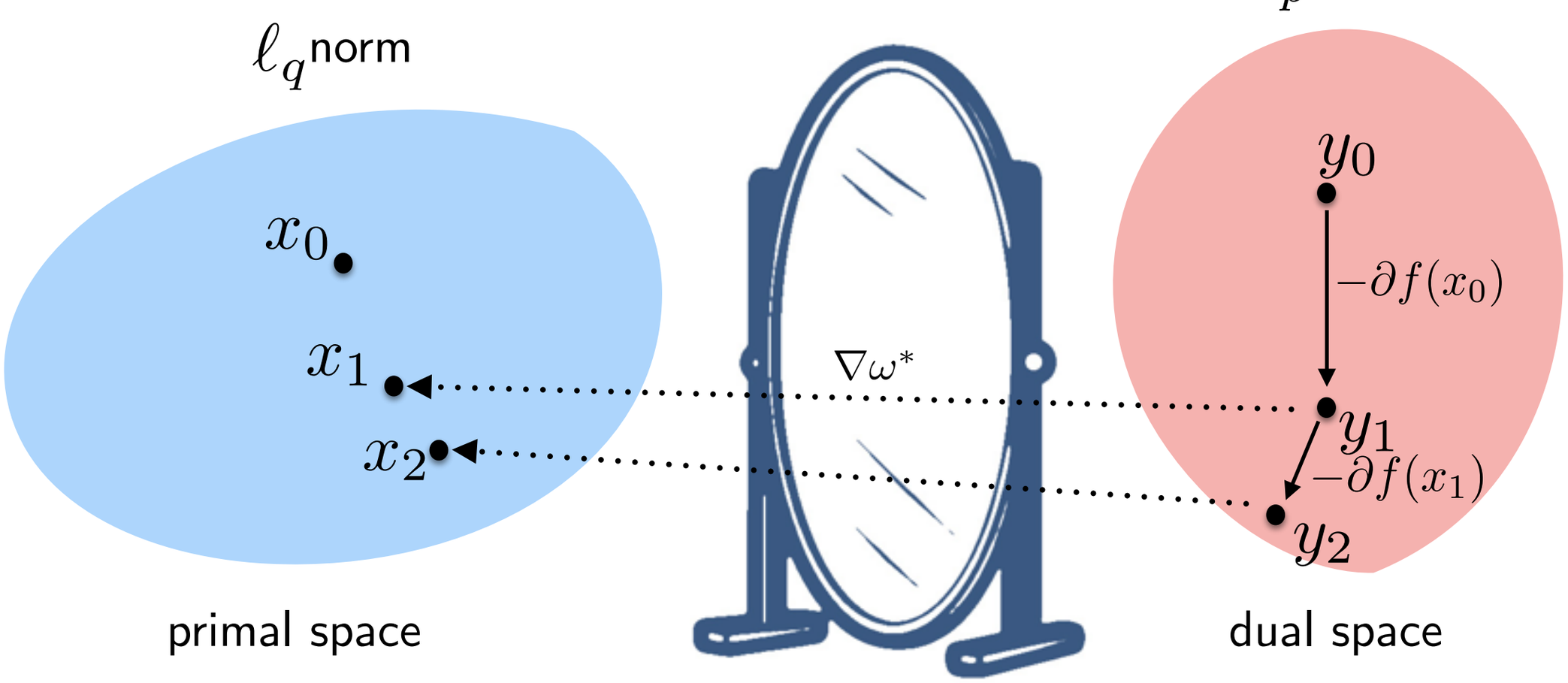


Mirror Descent

- Mirror Descent Iteration:

$$y_{t+1} = y_t - \eta_t \nabla f(x_t)$$

$$x_{t+1} = \nabla \omega^*(y_t)$$



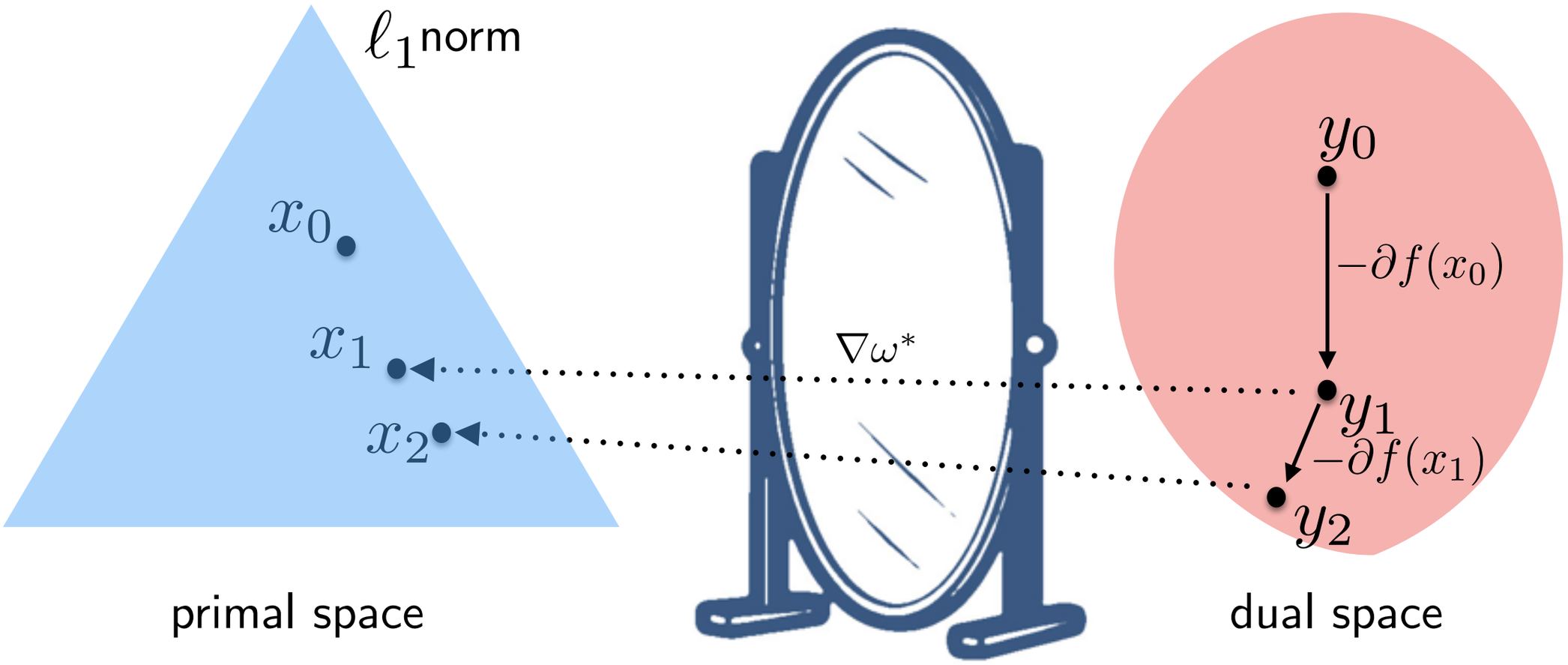
Multiplicative Weight Updates

- Mirror Descent Iteration:

$$y_{t+1} = y_t - \eta_t \nabla f(x_t)$$

$$x_{t+1} = \nabla \omega^*(y_t)$$

$$\nabla \omega^*(y)_i = \frac{e^{y_i}}{\sum_j e^{y_j}}$$



Mirror Descent for Carathéodory

- Mirror Descent Iteration:

$$y_{t+1} = y_t - \eta_t \nabla f(x_t)$$

$$x_{t+1} = \nabla \omega^*(y_t)$$

(template to be instantiated by ω^*)

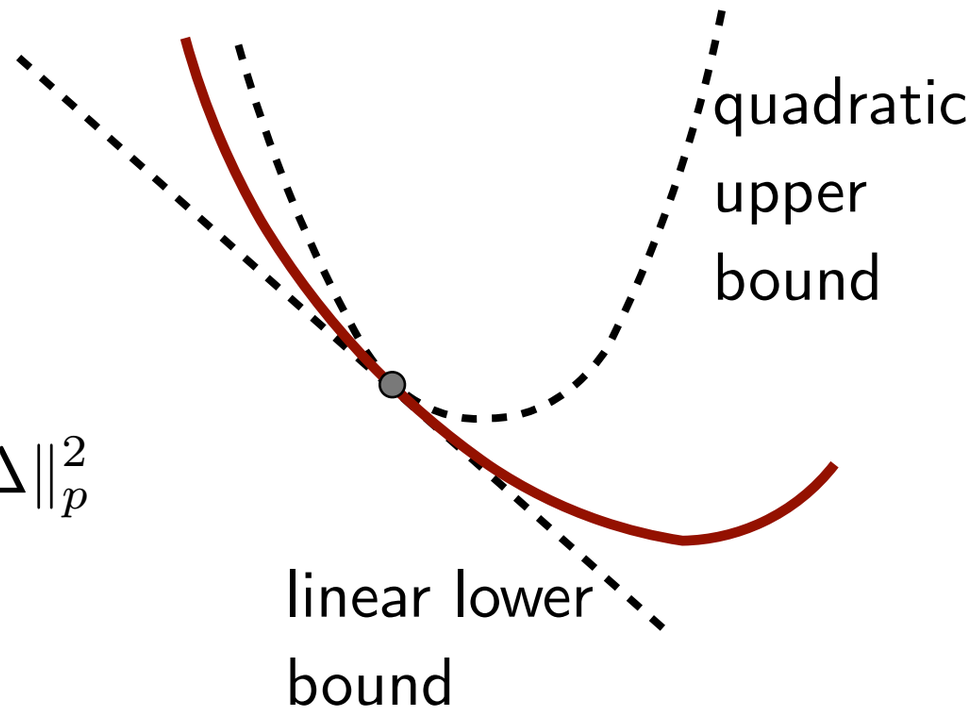
- Mirror Descent Guarantee:

$\frac{1}{T}\sum_t \nabla f(x_t)^\top (x_t - x) \leq \frac{RL}{\sqrt{T}}$

R = $\max_x D_\omega(x \| x_0)^{1/2}$
(distance function induced by ω^*)
L = ℓ_p-Lipschitz constant

- Requirements on ω^* :
  - $\nabla \omega^*(x) \in \mathcal{D}$
  - 1-smoothness in the ℓ_p-norm:

$$\omega^*(y + \Delta) \leq \omega^*(y) + \nabla \omega^*(y)^\top \Delta + \frac{1}{2}\|\Delta\|_p^2$$



Mirror Descent for Carathéodory

- Our job is reduced to design the right map: ω^*

$$\omega^*(y) = \begin{cases} \dfrac{1}{2(p-1)}\|y\|_p^2, & \|y\|_p \le 1 \\ \dfrac{1}{p-1}\left(\|y\|_p - \dfrac{1}{2}\right), & \|y\|_p > 1 \end{cases}$$

- **Final algorithm**:

$y_1 = 0 \quad z_1 = \nabla\omega^*(y_1)$

for t $= 1..O(p/\epsilon^2)$

$z_t = \nabla\omega^*(y_t)$

find v_i minimizing $v_i^\top z_t$ — most expensive step

$y_{t+1} = y_t + \eta \cdot (v_i - u)$

Overall $O(Np/\epsilon^2)$ running time.

Mirror Descent for Carathéodory

- Our job is reduced to design the right map: ω^*

$$\omega^*(y) = \begin{cases} \dfrac{1}{2(p-1)}\|y\|_p^2, & \|y\|_p \leq 1 \\ \dfrac{1}{p-1}\left(\|y\|_p - \dfrac{1}{2}\right), & \|y\|_p > 1 \end{cases}$$

- **Final algorithm**:

$y_1 = 0 \quad z_1 = \nabla\omega^*(y_1)$

for t $= 1..O(p/\epsilon^2)$

$z_t = \nabla\omega^*(y_t)$

find v_i minimizing $v_i^\top z_t$

$y_{t+1} = y_t + \eta \cdot (v_i - u)$

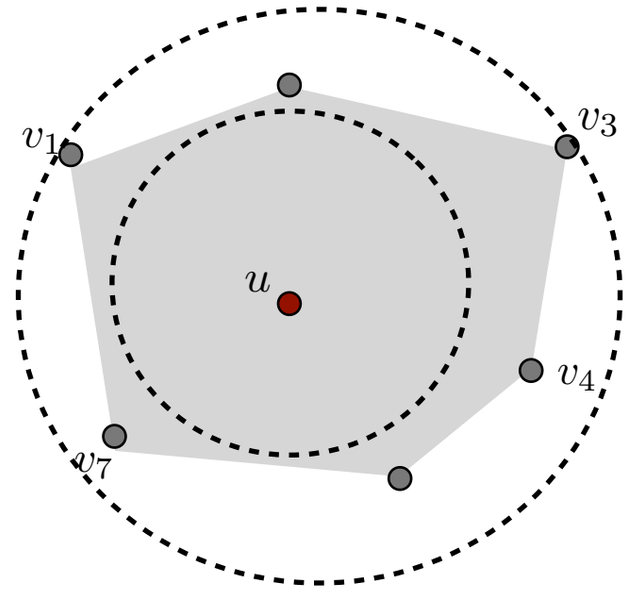
- **Nice feature**: no need to know a representation of

$$\{v_1, \ldots, v_n\}$$

it is enough to know how to optimize over this set: e.g., set of basis of a matroid, set of matchings, ...

Extensions and Applications

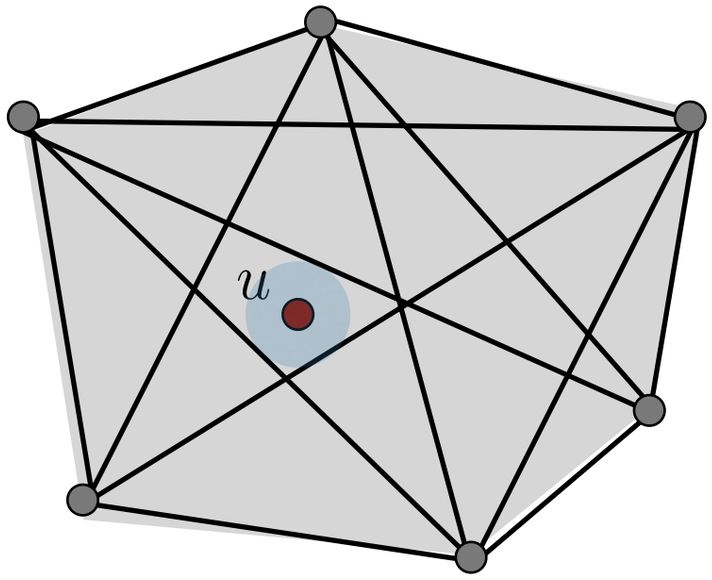
- Fat Caratheodory Theorem: if
$$B_p(u;r) \subseteq \text{hull}(V) \subseteq B_p(u;1)$$
then the approximate Caratheodory Theorem holds with $k = O\left(\frac{p}{r^2} \cdot \log \frac{r}{\epsilon}\right)$



- Rounding in polytopes.
- Training SVMs.
- Accelerating Fujishige's algorithms for submodular minimization.

Lower bound

- Fixed p and ϵ , then for sufficiently large n, there is an n x n matrix V with unit ℓ_p-columns and $x \in \Delta_n$ such that for all $x' \in \Delta, \|x'\|_0 \leq cp/\epsilon^2$, $\|Vx - Vx'\|_p > \epsilon$



- Based on an idea by [Klein, Young] to bound the # of iterations of Danzig-Wolfe methods in Linear Programming.

- Start with the saddle point formulation:

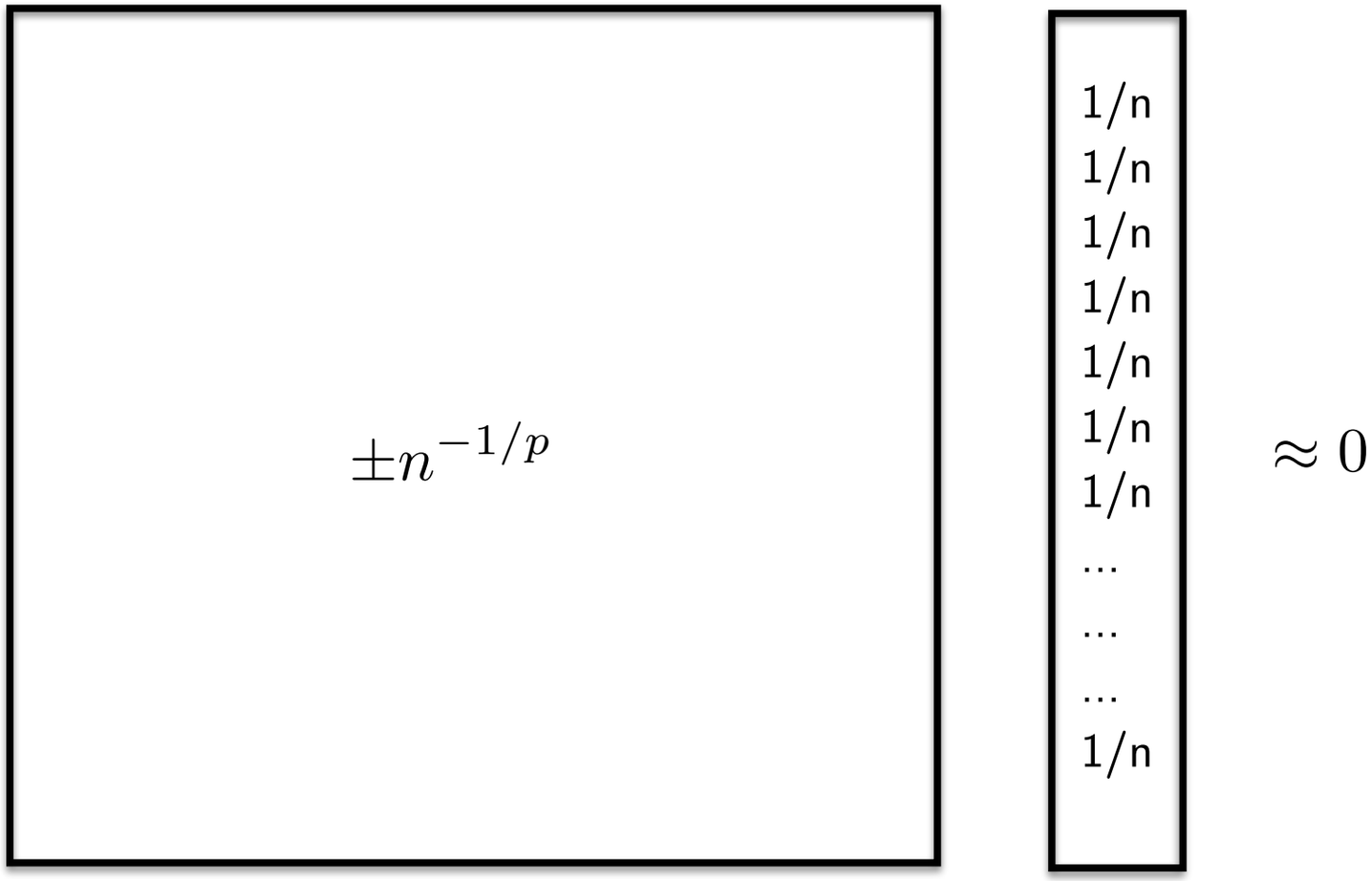
$$0 = \min_{x \in \Delta} \max_{y \in B_q(1)} y^\top (Vx - u)$$

We want to show that:

$$\min_{x \in \Delta, \|x\|_0 < k} \max_{y \in B_q(1)} y^\top (Vx - u) > \epsilon$$

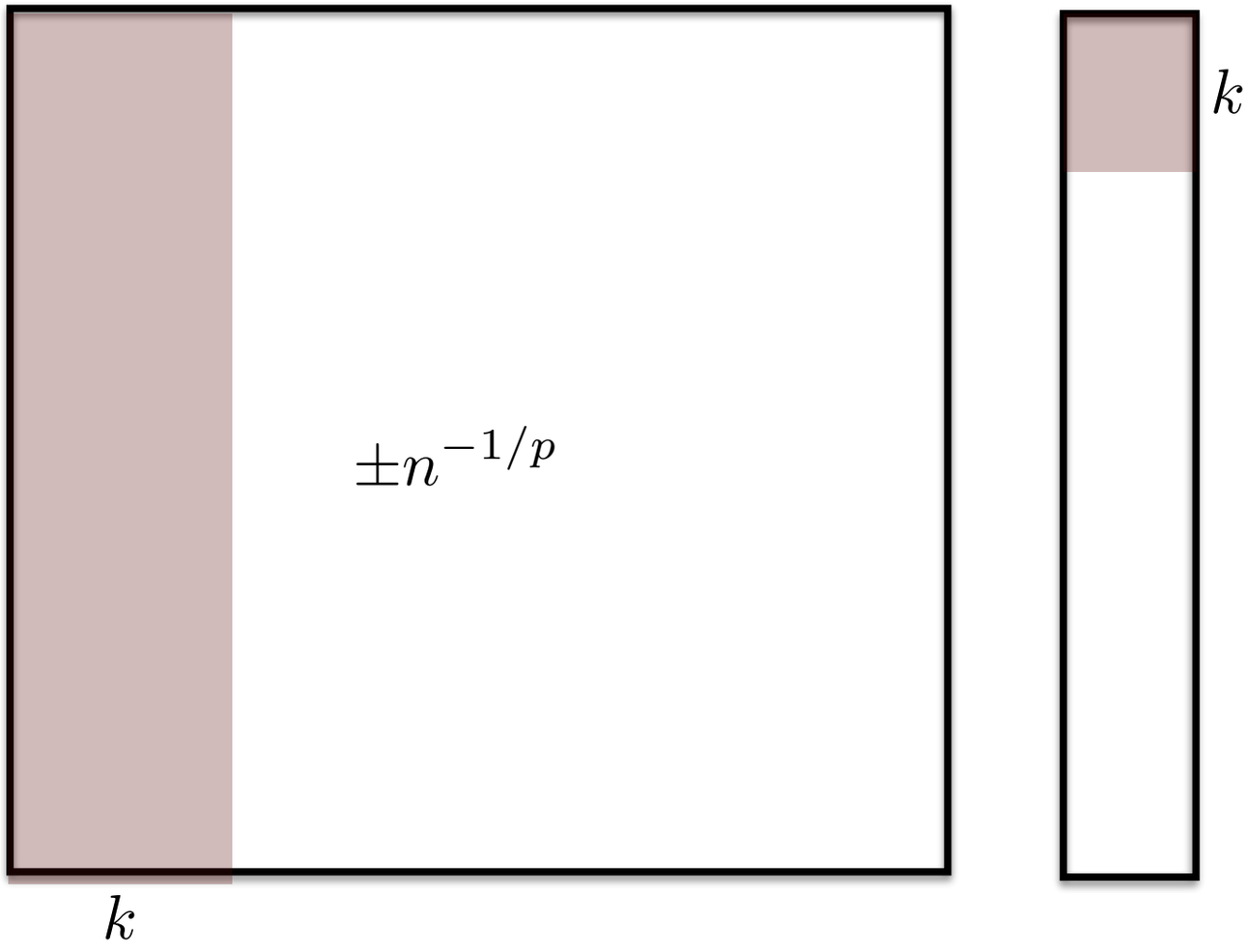
Lower bound

- Let V be a random n x n matrix of iid $\pm n^{-1/p}$ entries.
- Let $x = \mathbf{1}/n$ be the uniform vector, so that $Vx \approx 0$



Lower bound

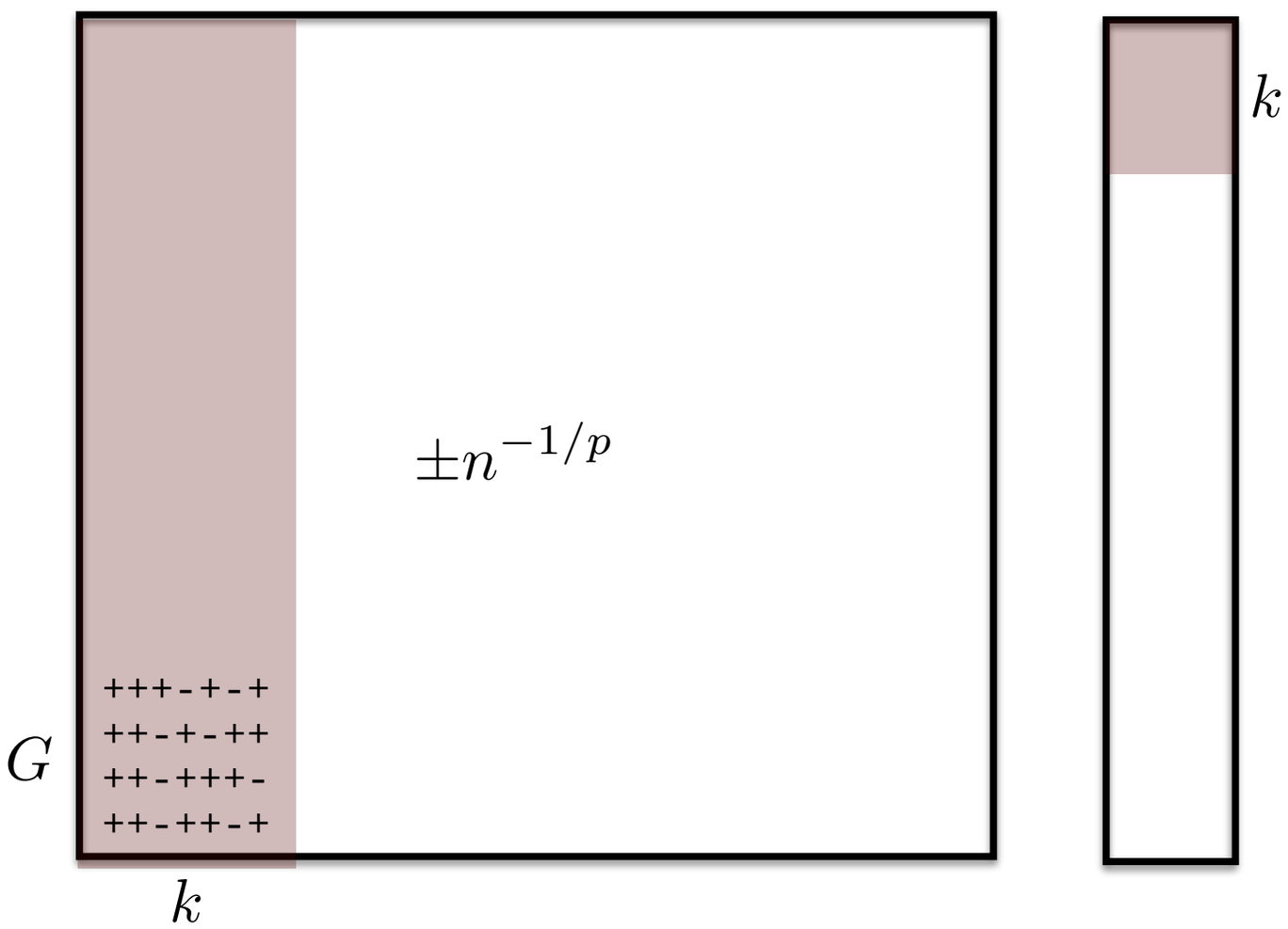
- Let V be a random n x n matrix of iid $\pm n^{-1/p}$ entries.
- Let $x = \mathbf{1}/n$ be the uniform vector, so that $Vx \approx 0$
- Assume $\mathrm{supp}(x) = S$ with $|S| = k$ constant independent of n.



Lower bound

- For each set S, there is (whp) a set G of very skewed rows.

- Set y to be uniform over G.



Lower bound

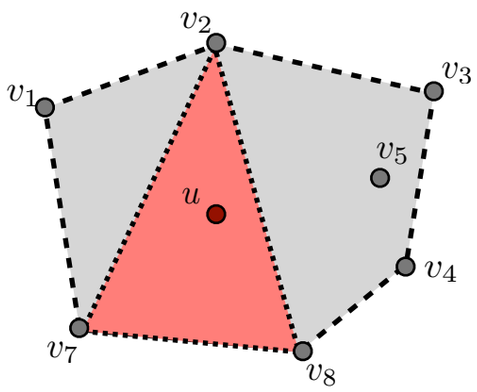
- For each set S, there is (whp) a set G of very skewed rows.

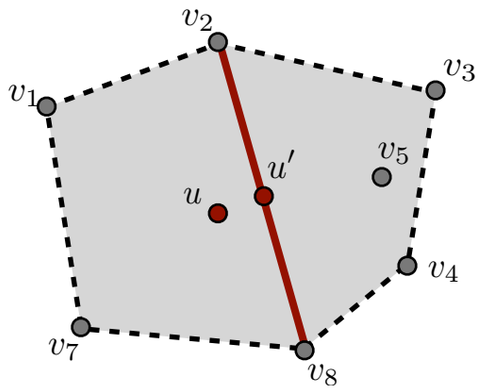
- For every set S, an adversary y can play an uniform distribution of those rows to force a large value of $y^\top V x$.

- Applying the appropriate concentration bounds, whp, the adversary can choose for every S and y to force the value to be $> \epsilon \left(\frac{r}{n}\right)^{1/p}$ for $r = Cn \exp(-ck\epsilon^2)$.

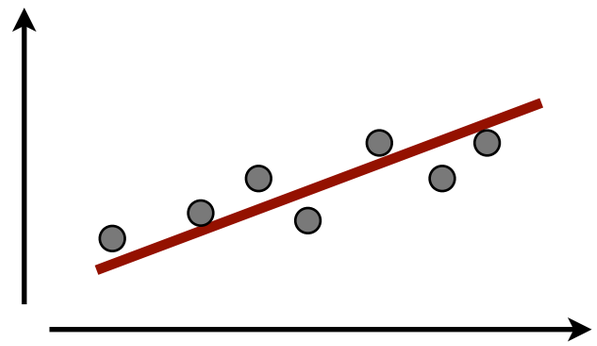
- So for $k < Cp/\epsilon^2$ the adversary can force a value of at least ϵ .

Sparsification via Optimization

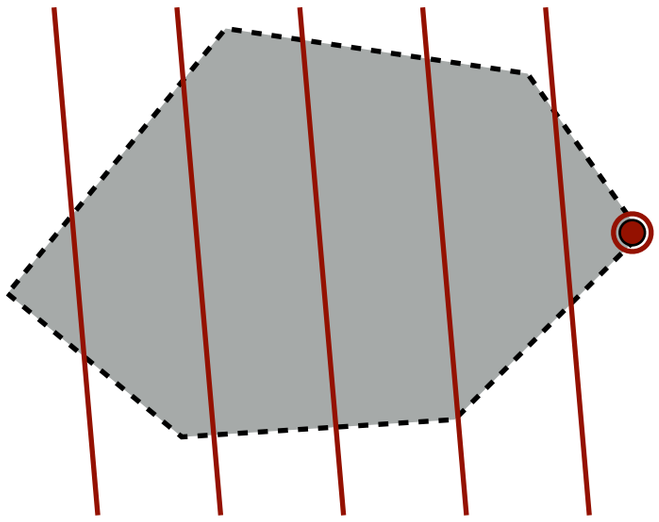




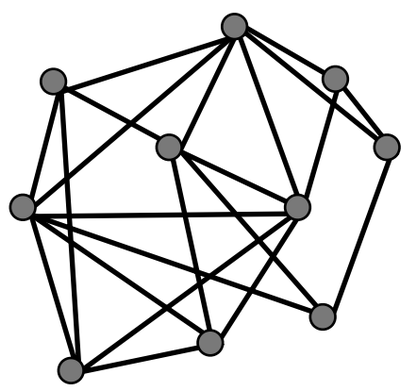
This paper

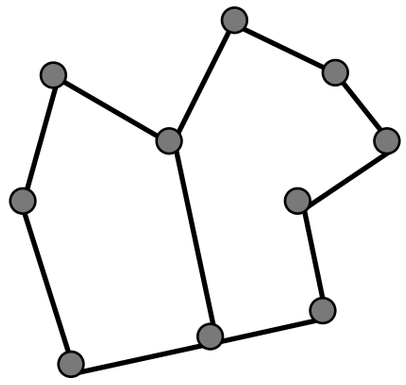


[Shalev-Shwartz, Srebro, Zhang]



[Plotkin, Shmoys, Tardos]





[Allen-Zhu, Orecchia, Liao]

Questions

- Chernoff bounds —> Optimization ?

- Other problems where concentration bound can be replaced by mirror descent.