MACHINE LEARNING DEPARTMENT

Introduction to Machine Learning

5. Optimization

Geoff Gordon and Alex Smola
Carnegie Mellon University

http://alex.smola.org/teaching/cmu2013-10-701
10-701x

Carnegie Mellon University

Class Scoreboard for homework1

10-701 Classification Contest!

0	NICKNAME	VERSION	TIME	CLASSIFICATION A
1	Unknown	52	2013-09-28 01:34:34	66.25%
2	data,data,data	45	2013-09-29 18:47:58	66.75%
3	ASD	52	2013-09-29 19:24:33	66.75%
4	fuzzyaxioms	21	2013-09-29 19:38:03	66.5%
5	skywalker	10	2013-09-29 23:40:02	65.75%
6	(^.^)~c{-_-")	14	2013-09-27 19:04:37	64%
7	dloates	38	2013-09-29 19:21:05	64%
8	siyuano	48	2013-09-29 23:46:58	64.25%
9	Barack Obama	56	2013-09-30 02:19:44	64.25%
10	shock	40	2013-09-28 15:59:04	63.75%

Optimization

- Basic Techniques
  - Gradient descent
  - Newton's method
- Constrained Convex Optimization
  - Properties
  - Lagrange function
  - Wolfe dual
- Batch methods
  - Distributed subgradient
  - Bundle methods
- Online methods
  - Unconstrained subgradient
  - Gradient projections
  - Parallel optimization

Why

Parameter Estimation

- Maximum a Posteriori with Gaussian Prior

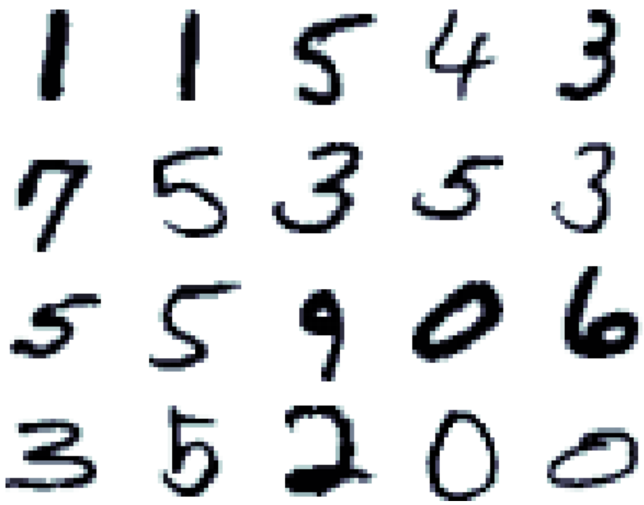
$$-\log p(\theta|X) = \frac{1}{2\sigma^2} \|\theta\|^2 + \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle + \text{const.}$$

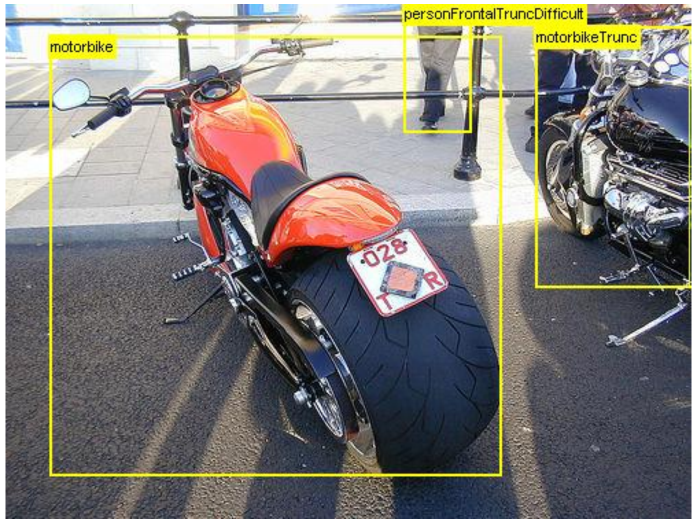
prior

data

- We have lots of data
  - Does not fit on single machine
  - Bandwidth constraints
  - May grow in real time
- Regularized Risk Minimization yields similar problems (more on this in a later lecture)

Batch and Online



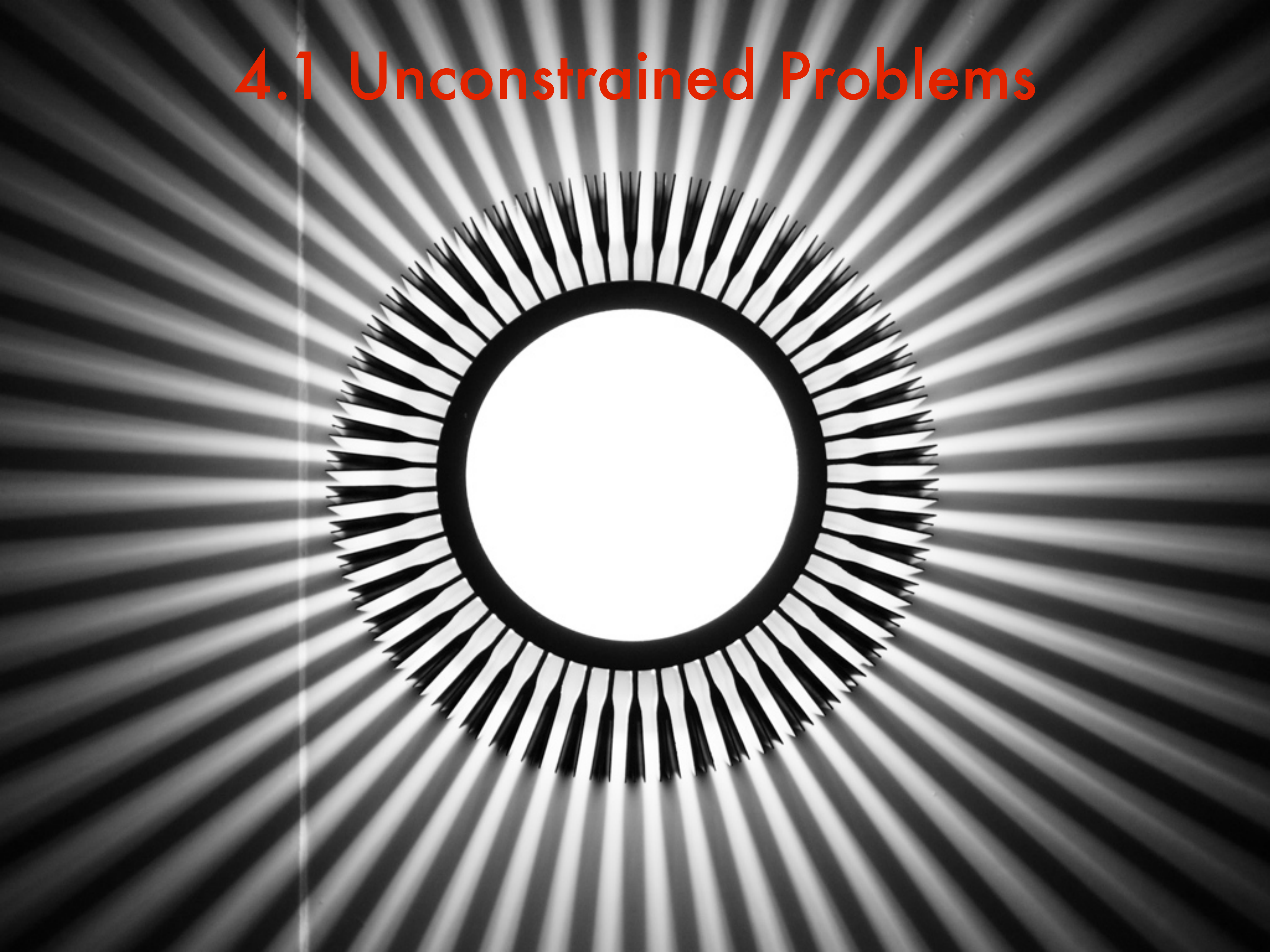


- Batch
  - Very large dataset available
  - Require parameter only at the end
    - optical character recognition
    - speech recognition
    - image annotation / categorization
    - machine translation
- Online
  - Spam filtering
  - Computational advertising
  - Content recommendation / collaborative filtering

Many parameters

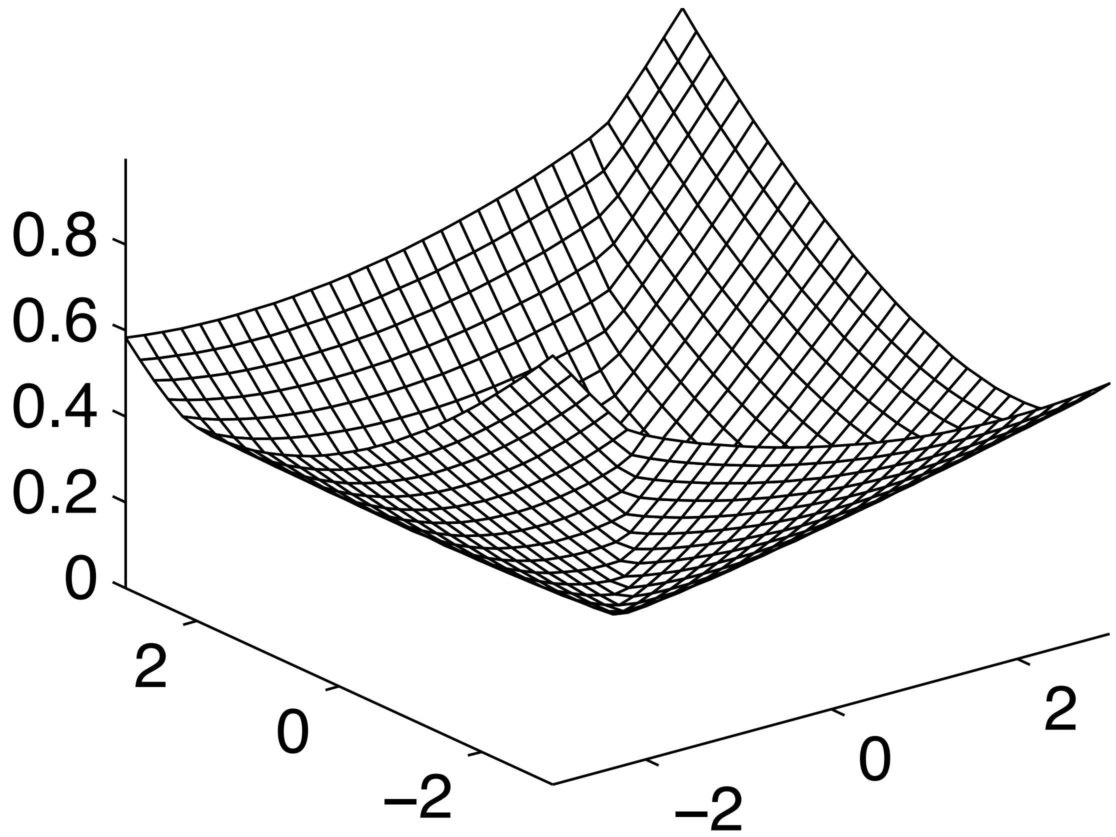
- 100 million to 1 Billion users
  Personalized content provision - impossible to adjust all parameters by heuristic/manually
- 1,000-10,000 computers
  Cannot exchange all data between machines, Distributed optimization, multicore
- Large networks
  Nontrivial parameter dependence structure

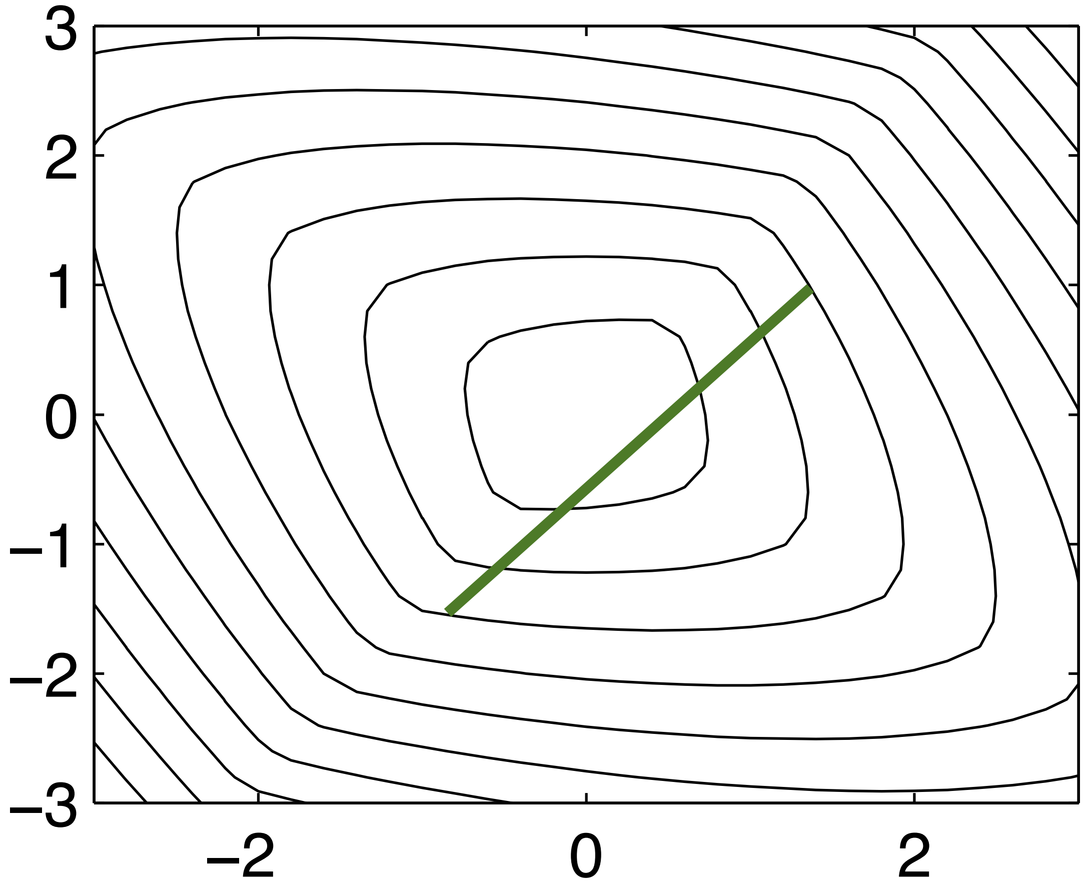
4.1 Unconstrained Problems



Convexity 101

Convexity 101





- **Convex set**

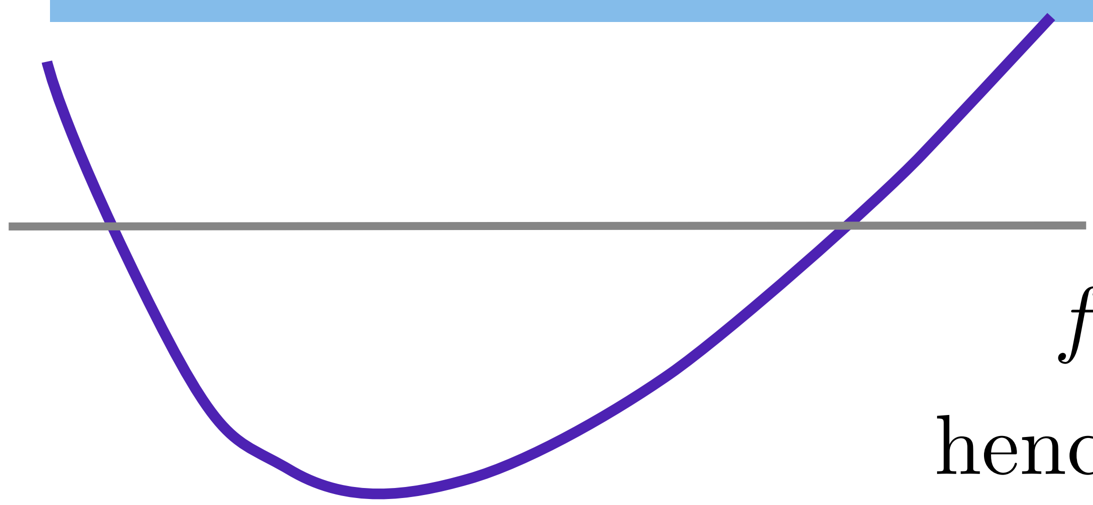
  For $x, x' \in X$ it follows that $\lambda x + (1-\lambda)x' \in X$ for $\lambda \in [0,1]$

- **Convex function**

$$\lambda\lambda f(x) + (1-\lambda) f(x') \geq f(\lambda x + (1-\lambda)x') \text{ for } \lambda \in [0,1]$$

Convexity 101

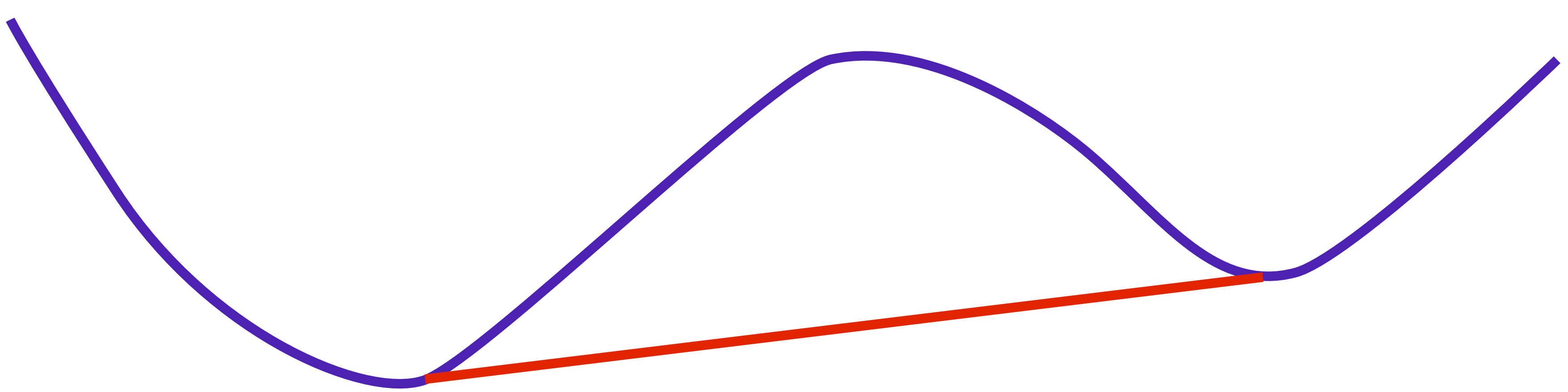
- Below-set of convex function is convex



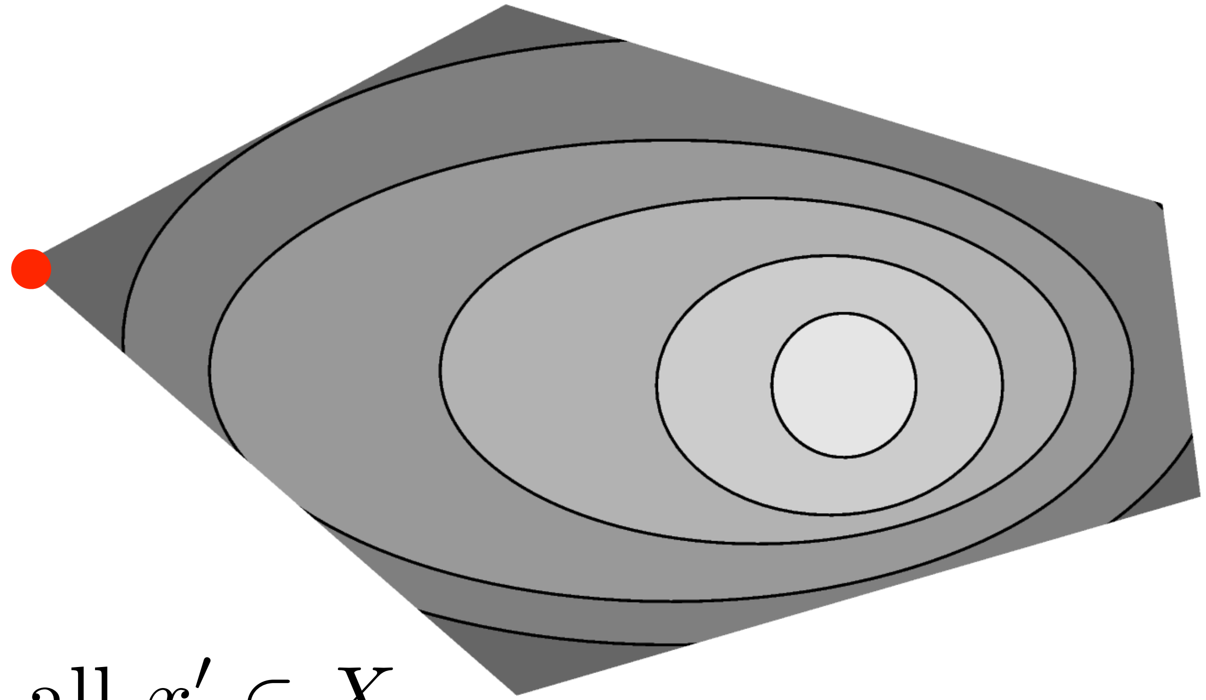
$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x$$

$$\text{hence } \lambda x + (1-\lambda)x' \in X \text{ for } x, x' \in X$$

- Convex functions don’t have local minima
  Proof by contradiction - linear interpolation breaks local minimum condition



Convexity 101



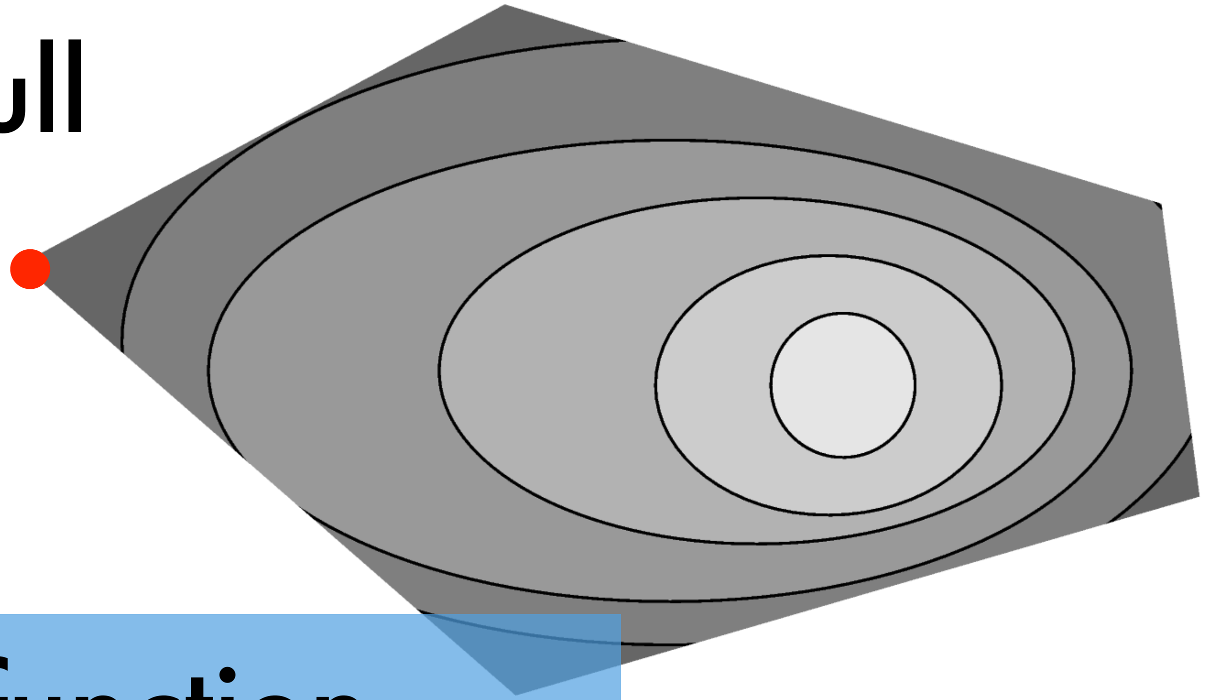
- **Vertex of a convex set**
  Point which cannot be extrapolated within convex set
  $\lambda x + (1-\lambda)x' \notin X$ for $\lambda > 1$ for all $x' \in X$
- **Convex hull**
  $$\operatorname{co} X := \left\{ \bar{x} \,\middle|\, \bar{x} = \sum_{i=1}^{n} \alpha_i x_i \text{ where } n \in \mathbb{N}, \alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i \leq 1 \right\}$$
- Convex hull of set is a convex set (proof trivial)

Convexity 101

- Supremum on convex hull

$$\sup_{x \in X} f(x) = \sup_{x \in \mathrm{co} X} f(x)$$

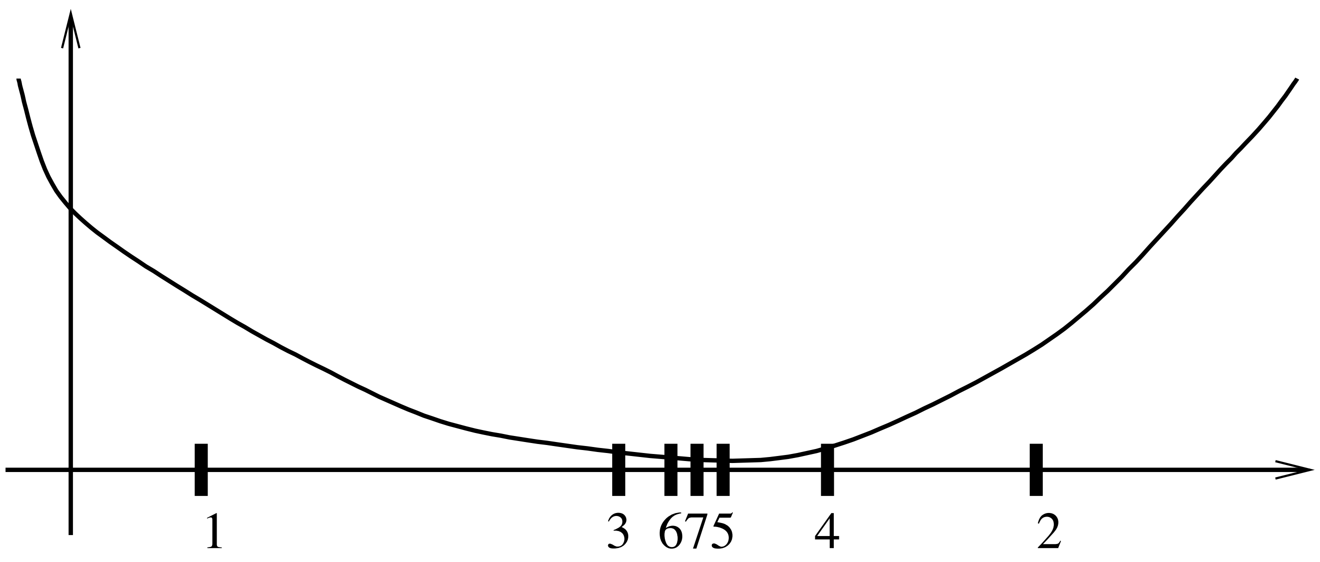
Proof by contradiction



- Maximum over convex function on convex set is obtained on vertex
  - Assume that maximum inside line segment
  - Then function cannot be convex
  - Hence it must be on vertex

Gradient descent

One dimensional problems

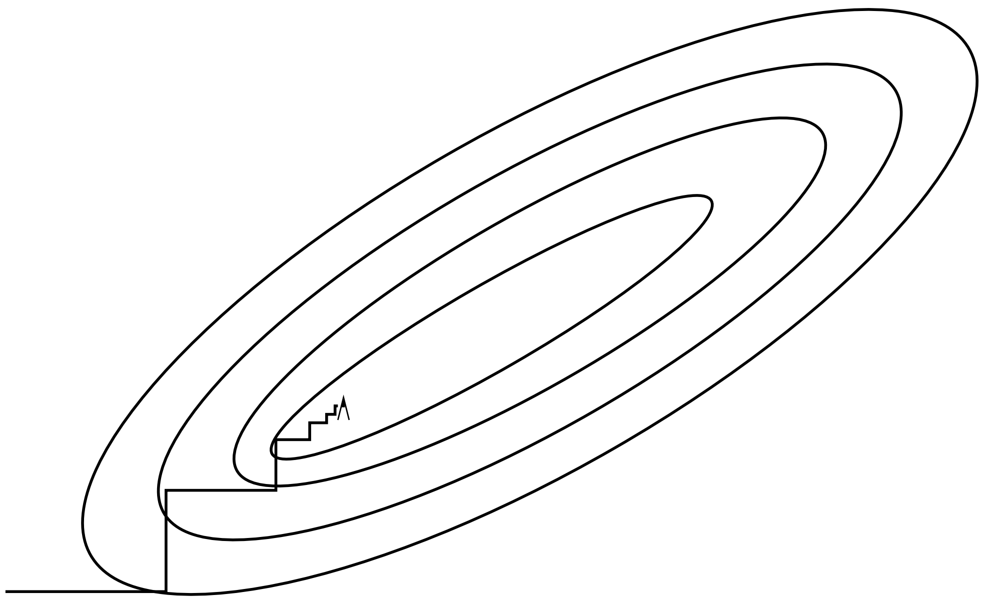


Require: $a, b,$ Precision ϵ
Set $A = a, B = b$
repeat
if $f'\left(\frac{A+B}{2}\right) > 0$ **then**
$B = \frac{A+B}{2}$
else
$A = \frac{A+B}{2}$
end if
until $(B - A)\min(|f'(A)|, |f'(B)|) \leq \epsilon$
Output: $x = \frac{A+B}{2}$

solution on the left

- Key Idea
  - For differentiable f search for x with f'(x) = 0
  - Interval bisection (derivative is monotonic)
  - Need log (A-B) - log ε to converge
- Can be extended to nondifferentiable problems (exploit convexity in upper bound and keep 5 points)

Gradient descent



- Key idea
  - Gradient points into descent direction
  - Locally gradient is good approximation of objective function
- GD with Line Search
  - Get descent direction
  - Unconstrained line search
  - Exponential convergence for strongly convex objective

given a starting point $x \in \textbf{dom}\, f$.

repeat

1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

Convergence Analysis

- Strongly convex function

$$f(y) \geq f(x) + \langle y - x, \partial_x f(x) \rangle + \frac{m}{2} \left\| y - x \right\|^2$$

- Progress guarantees (minimum x*)

$$f(x) - f(x^*) \geq \frac{m}{2} \left\| x - x^* \right\|^2$$

- Lower bound on the minimum (set y= x*)

$$\begin{aligned} f(x) - f(x^*) &\leq \langle x - x^*, \partial_x f(x) \rangle - \frac{m}{2} \left\| x^* - x \right\|^2 \\ &\leq \sup_y \langle x - y, \partial_x f(x) \rangle - \frac{m}{2} \left\| y - x \right\|^2 \\ &= \frac{1}{2m} \left\| \partial_x f(x) \right\|^2 \end{aligned}$$

Convergence Analysis

- **Bounded Hessian**

$$f(y) \leq f(x) + \langle y - x, \partial_x f(x) \rangle + \frac{M}{2} \|y - x\|^2$$

$$\implies f(x + tg_x) \leq f(x) - t\|g_x\|^2 + \frac{M}{2} t^2 \|g_x\|^2$$

$$\leq f(x) - \frac{1}{2M} \|g_x\|^2$$

Using strong convexity

$$\implies f(x + tg_x) - f(x^*) \leq f(x) - f(x^*) - \frac{1}{2M} \|g_x\|^2$$

$$\leq f(x) - f(x^*) \left[1 - \frac{m}{M}\right]$$

- **Iteration bound** $\frac{M}{m} \log \frac{f(x) - f(x^*)}{\epsilon}$

Newton’s Method

Isaac Newton

Newton Method

- Convex objective function f
- Nonnegative second derivative

$$\partial_x^2 f(x) \succeq 0$$

- Taylor expansion

$$f(x+\delta) = f(x) + \langle \delta, \partial_x f(x) \rangle + \frac{1}{2}\delta^\top \partial_x^2 f(x) \delta + O(\delta^3)$$

gradient

Hessian

- Minimize approximation & iterate til converged

$$x \leftarrow x - \left[\partial_x^2 f(x)\right]^{-1} \partial_x f(x)$$

Convergence Analysis

- There exists a region around optimality where Newton's method converges quadratically if f is twice continuously differentiable
- For some region around x* gradient is well approximated by Taylor expansion

$$\left\|\partial_x f(x^*) - \partial_x f(x) - \left\langle x^* - x, \partial_x^2 f(x) \right\rangle\right\| \leq \gamma \left\|x^* - x\right\|^2$$

- Expand Newton update

$$\begin{aligned}
\|x_{n+1} - x^*\| &= \left\| x_n - x^* - \left[\partial_x^2 f(x_n)\right]^{-1} \left[\partial_x f(x_n) - \partial_x f(x^*)\right] \right\| \\
&= \left\| \left[\partial_x^2 f(x_n)\right]^{-1} \left[\partial_x^f (x_n)[x_n - x^*] - \partial_x f(x_n) + \partial_x f(x^*)\right] \right\| \\
&\leq \gamma \left\| \left[\partial_x^2 f(x_n)\right]^{-1} \right\| \|x_n - x^*\|^2
\end{aligned}$$

Convergence Analysis

- Two convergence regimes
  - As slow as gradient descent outside the region where Taylor expansion is good

$$\left\|\partial_x f(x^*) - \partial_x f(x) - \left\langle x^* - x, \partial_x^2 f(x)\right\rangle\right\| \leq \gamma \left\|x^* - x\right\|^2$$

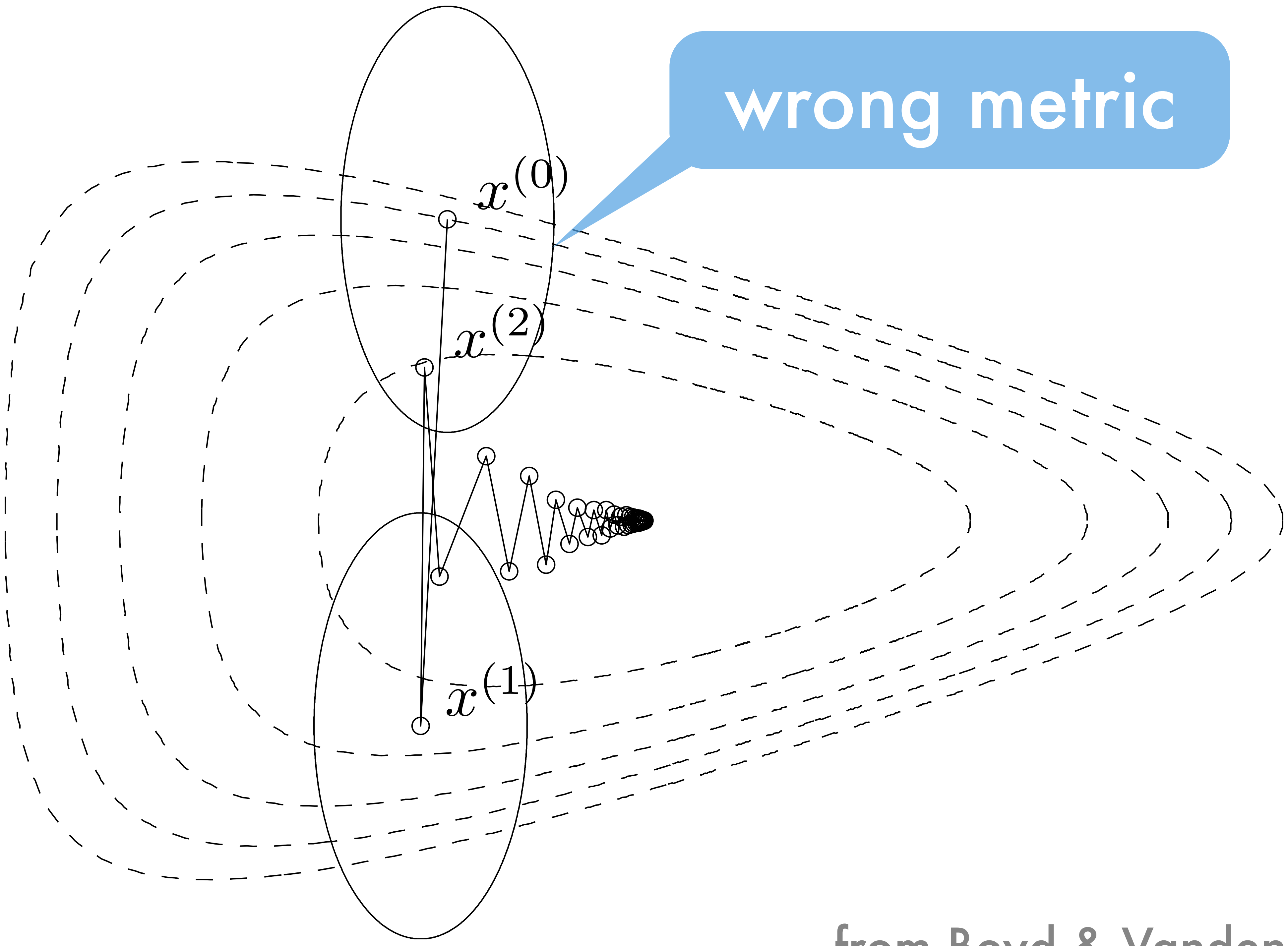
  - Quadratic convergence once the bound holds

$$\|x_{n+1} - x^*\| \leq \gamma \left\|\left[\partial_x^2 f(x_n)\right]^{-1}\right\| \|x_n - x^*\|^2$$

  - Newton method is affine invariant (proof by chain rule)

See Boyd and Vandenberghe, Chapter 9.5 for much more

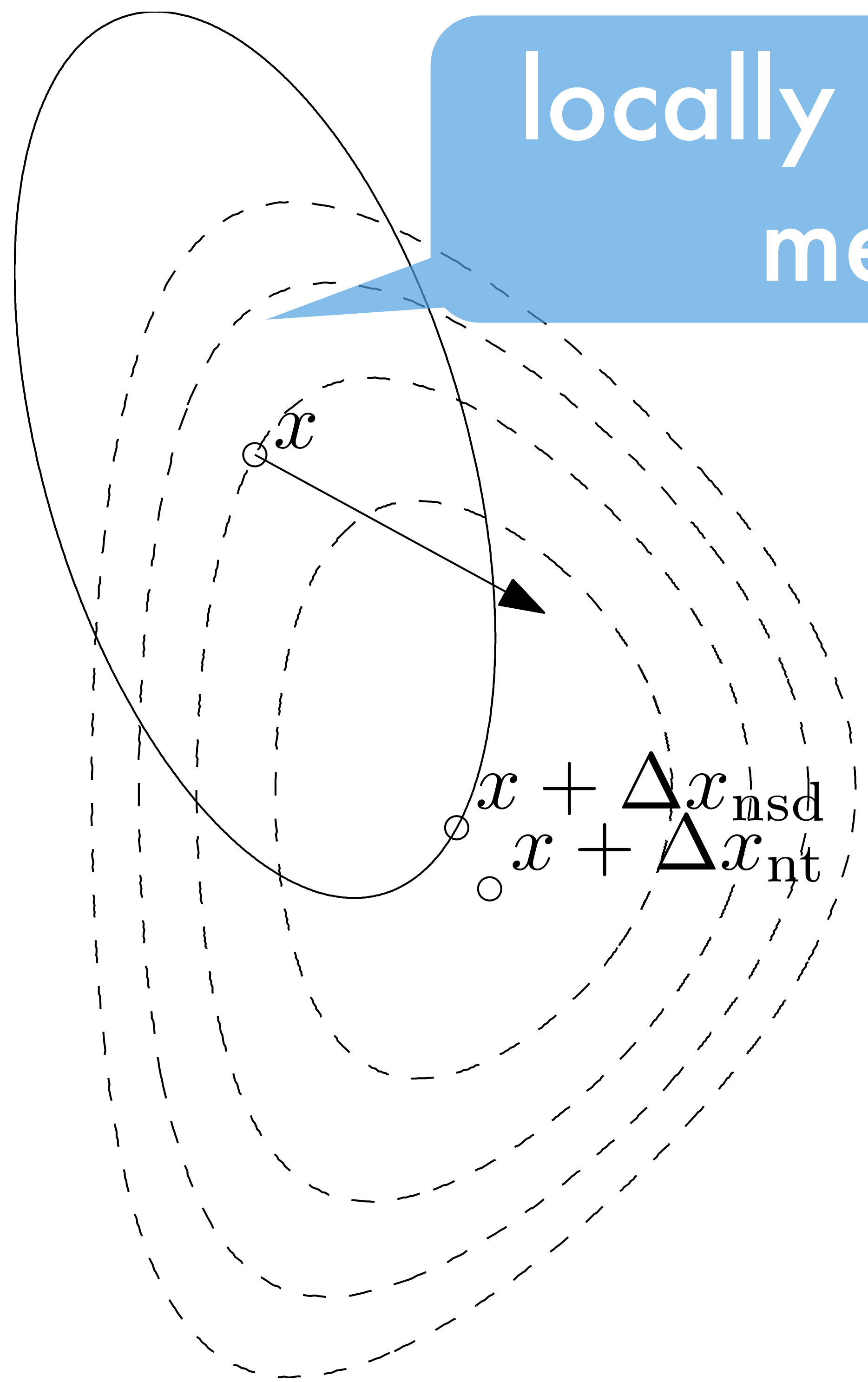
Newton method rescales space



wrong metric

from Boyd & Vandenberghe

Newton method rescales space



locally adaptive metric

Boyd & Vandenberghe

Parallel Newton Method

- Good rate of convergence
- Few passes through data needed
- Parallel aggregation of gradient and Hessian
- Gradient requires $O(d)$ data
- Hessian requires $O(d^2)$ data
- Update step is $O(d^3)$ & nontrivial to parallelize
- Use it only for low dimensional problems

BFGS algorithm
Broyden-Fletcher-Goldfarb-Shanno

Basic Idea

- Newton-like method to compute descent direction

$$\delta_i = B_i^{-1} \partial_x f(x_{i-1})$$

- Line search on f in direction

$$x_{i+1} = x_i - \alpha_i \delta_i$$

- Update B with rank 2 matrix

$$B_{i+1} = B_i + u_i u_i^\top + v_i v_i^\top$$

- Require that Quasi-Newton condition holds

$$B_{i+1}(x_{i+1} - x_i) = \partial_x f(x_{i+1}) - \partial_x f(x_i)$$

$$B_{i+1} = B_i + \frac{g_i g_i^\top}{\alpha_i \delta_i^\top g_i} - \frac{B_i \delta_i \delta_i^\top B_i}{\delta_i^\top B_i \delta_i}$$

Properties

- Simple rank 2 update for B
- Use matrix inversion lemma to update inverse
- Memory-limited versions L-BFGS
- Use toolbox if possible (TAO, MATLAB)
(typically slower if you implement it yourself)
- Works well for nonlinear nonconvex objectives
(often even for nonsmooth objectives)

4.2 Constrained Convex Problems

Basic Convexity

Constrained Convex Minimization

- Optimization problem

$$\underset{x}{\text{minimize}}\ f(x)$$

$$\text{subject to } c_i(x) \le 0 \text{ for}$$

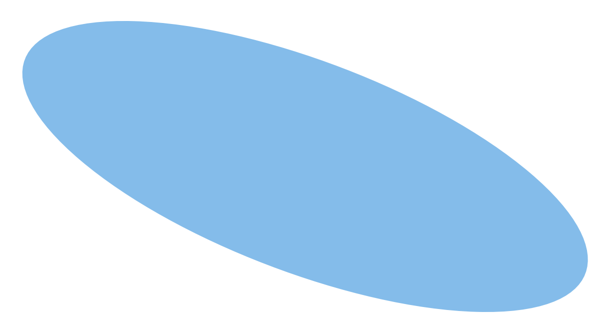
Equality is special case
Why?

- Common constraints
  - linear inequality constraints

$$\langle w_i, x\rangle + b_i \le 0$$

  - quadratic cone constraints

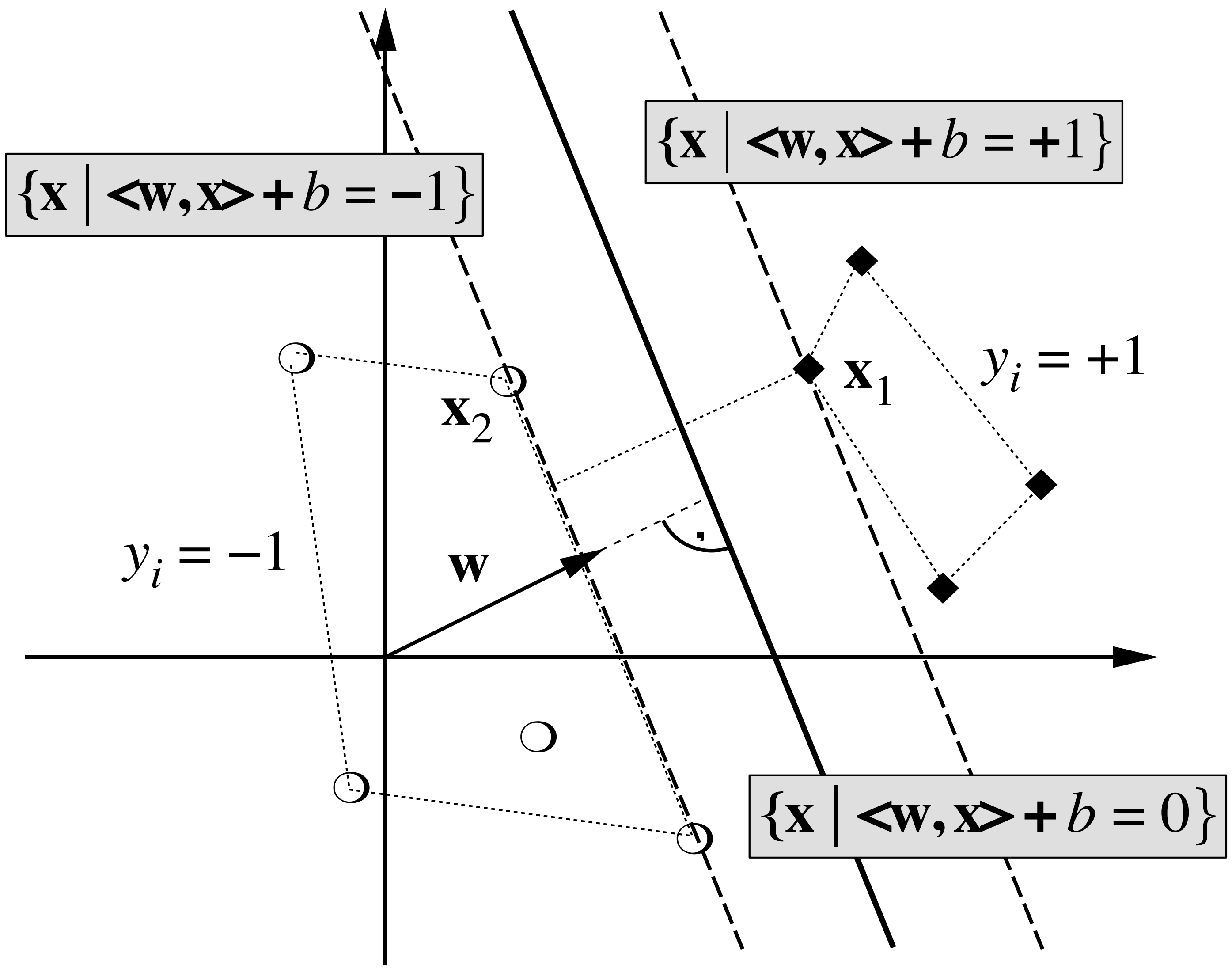
$$x^\top Q x + b^\top x \le c \text{ with } Q \succeq 0$$



  - semidefinite constraints

$$M \succeq 0 \text{ or } M_0 + \sum_i x_i M_i \succeq 0$$

Example - Support Vectors



$$\langle w, x_1 \rangle + b = 1$$

$$\langle w, x_2 \rangle + b = -1$$

$$\text{hence} \quad \langle w, x_1 - x_2 \rangle + b = 2$$

$$\text{hence} \quad \left\langle \frac{w}{\|w\|}, x_1 - x_2 \right\rangle = \frac{2}{\|w\|}$$

margin

$$\underset{w,b}{\text{minimize}} \; \frac{1}{2} \|w\|^2 \;\; \text{subject to } y_i \left[\langle w, x_i \rangle + b\right] \geq 1$$

Lagrange Multipliers

- Lagrange function

$$L(x, \alpha) := f(x) + \sum_{i=1}^{n} \alpha_i c_i(x) \text{ where } \alpha_i \geq 0$$

- Saddlepoint Condition
If there are x* and nonnegative α* such that

$$L(x^*, \alpha) \leq L(x^*, \alpha^*) \leq L(x, \alpha^*)$$

then x* is an optimal solution to the constrained optimization problem

Proof

$$L(x^*, \alpha) \leq L(x^*, \alpha^*) \leq L(x, \alpha^*)$$

- From first inequality we see that x* is feasible

$$(\alpha_i - \alpha_i^*) c_i(x^*) \leq 0 \text{ for all } \alpha_i \geq 0$$

- Setting some $\alpha_i = 0$ yields KKT conditions

$$\alpha_i^* c_i(x^*) = 0$$

- Consequently we have

$$L(x^*, \alpha^*) = f(x^*) \leq L(x, \alpha^*) = f(x) + \sum_i \alpha_i^* c_i(x) \leq f(x)$$

This proves optimality

Constraint gymnastics (all three conditions are equivalent)

- Slater's condition
  There exists some x such that for all i
  $$c_i(x) < 0$$
- Karlin's condition
  For all nonnegative α there exists some x such that
  $$\sum_i \alpha_i c_i(x) \leq 0$$
- Strict constraint qualification
  The feasible region contains at least two distinct elements and there exists an x in X such that all $c_i(x)$ are strictly convex at x with respect to X

Necessary Kuhn-Tucker Conditions

- Assume optimization problem
  - satisfies the constraint qualifications
  - has convex differentiable objective + constraints
- Then the KKT conditions are necessary & sufficient

$$\partial_x L(x^*, \alpha^*) = \partial_x f(x^*) + \sum_i \alpha_i^* \partial_x c_i(x^*) \qquad = 0 \text{ (Saddlepoint in } x^*\text{)}$$

$$\partial_{\alpha_i} L(x^*, \alpha^*) = c_i(x^*) \qquad \leq 0 \text{ (Saddlepoint in } \alpha^*\text{)}$$

$$\sum_i \alpha_i^* c_i(x^*) \qquad = 0 \text{ (Vanishing KKT-gap)}$$

Yields algorithm for solving optimization problems
Solve for saddlepoint and KKT conditions

Proof

$$
\begin{aligned}
f(x) - f(x^*) &\geq \left[\partial_x f(x^*)\right]^\top (x - x^*) && \text{(by convexity)} \\
&= -\sum_i \alpha_i^* \left[\partial_x c_i(x^*)\right]^\top (x - x^*) && \text{(by Saddlepoint in } x^*\text{)} \\
&\geq -\sum_i \alpha_i^* (c_i(x) - c_i(x^*)) && \text{(by convexity)} \\
&= \sum_i \alpha_i^* c_i(x) && \text{(by vanishing KKT gap)} \\
&\geq 0
\end{aligned}
$$

Linear and Quadratic Programs

Linear Programs

- **Objective**

$$\underset{x}{\text{minimize}}\, c^\top x \text{ subject to } Ax + d \leq 0$$

- **Lagrange function**

$$L(x, \alpha) = c^\top x + \alpha^\top (Ax + d)$$

- **Optimality conditions**

$$\partial_x L(x, \alpha) = A^\top \alpha + c = 0$$
$$\partial_\alpha L(x, \alpha) = Ax + d \leq 0$$
$$0 = \alpha^\top (Ax + d)$$
$$0 \leq \alpha$$

plug into L

- **Dual problem**

$$\underset{i}{\text{maximize}}\, d^\top \alpha \text{ subject to } A^\top \alpha + c = 0 \text{ and } \alpha \geq 0$$

Linear Programs

- Primal

$$\underset{x}{\text{minimize}}\, c^\top x \text{ subject to } Ax + d \leq 0$$

- Dual

$$\underset{i}{\text{maximize}}\, d^\top \alpha \text{ subject to } A^\top \alpha + c = 0 \text{ and } \alpha \geq 0$$

- Free variables become equality constraints
- Equality constraints become free variables
- Inequalities become inequalities
- Dual of dual is primal

Quadratic Programs

- **Objective**

$$\underset{x}{\text{minimize}} \; \frac{1}{2} x^\top Q x + c^\top x \text{ subject to } Ax + d \leq 0$$

- **Lagrange function**

$$L(x, \alpha) = \frac{1}{2} x^\top Q x + c^\top x + \alpha^\top (Ax + d)$$

- **Optimality conditions**

$$\begin{aligned} \partial_x L(x, \alpha) &= Qx + A^\top \alpha + c = 0 \\ \partial_\alpha L(x, \alpha) &= Ax + d \leq 0 \\ 0 &= \alpha^\top (Ax + d) \\ 0 &\leq \alpha \end{aligned}$$

plug into L

Quadratic Program

- Eliminating x from the Lagrangian via

$$Qx + A^\top \alpha + c = 0$$

- Lagrange function

$$\begin{aligned}
L(x, \alpha) &= \frac{1}{2} x^\top Q x + c^\top x + \alpha^\top (Ax + d) \\
&= -\frac{1}{2} x^\top Q x + \alpha^\top d \\
&= -\frac{1}{2} (A^\top \alpha + c)^\top Q^{-1} (A^\top \alpha + c) + \alpha^\top d \\
&= -\frac{1}{2} \alpha^\top A Q^{-1} A^\top \alpha + \alpha^\top \left[d - A Q^{-1} c \right] - \frac{1}{2} c^\top Q^{-1} c
\end{aligned}$$

$$\text{subject to } \alpha \geq 0$$

dual

Quadratic Programs

- Primal

$$\underset{x}{\text{minimize}}\ \frac{1}{2}x^\top Q x + c^\top x \text{ subject to } Ax + d \leq 0$$

- Dual

$$\underset{\alpha}{\text{minimize}}\ \frac{1}{2}\alpha^\top A Q^{-1} A^\top \alpha + \alpha^\top \left[A Q^{-1} c - d\right] \text{ subject to } \alpha \geq 0$$

  - Dual constraints are simpler
  - Possibly many fewer variables
  - Dual of dual is not (always) primal (e.g. in SVMs x is in a Hilbert Space)

Bundle Methods

simple parallelization

Some optimization problems

- **Density estimation**

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} -\log p(x_i|\theta) - \log p(\theta)$$

$$\text{equivalently} \quad \underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \left[g(\theta) - \langle \phi(x_i), \theta \rangle\right] + \frac{1}{2\sigma^2} \|\theta\|^2$$

- **Penalized regression**

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} l\left(y_i - \langle \phi(x_i), \theta \rangle\right) + \frac{1}{2\sigma^2} \|\theta\|^2$$

e.g. squared loss

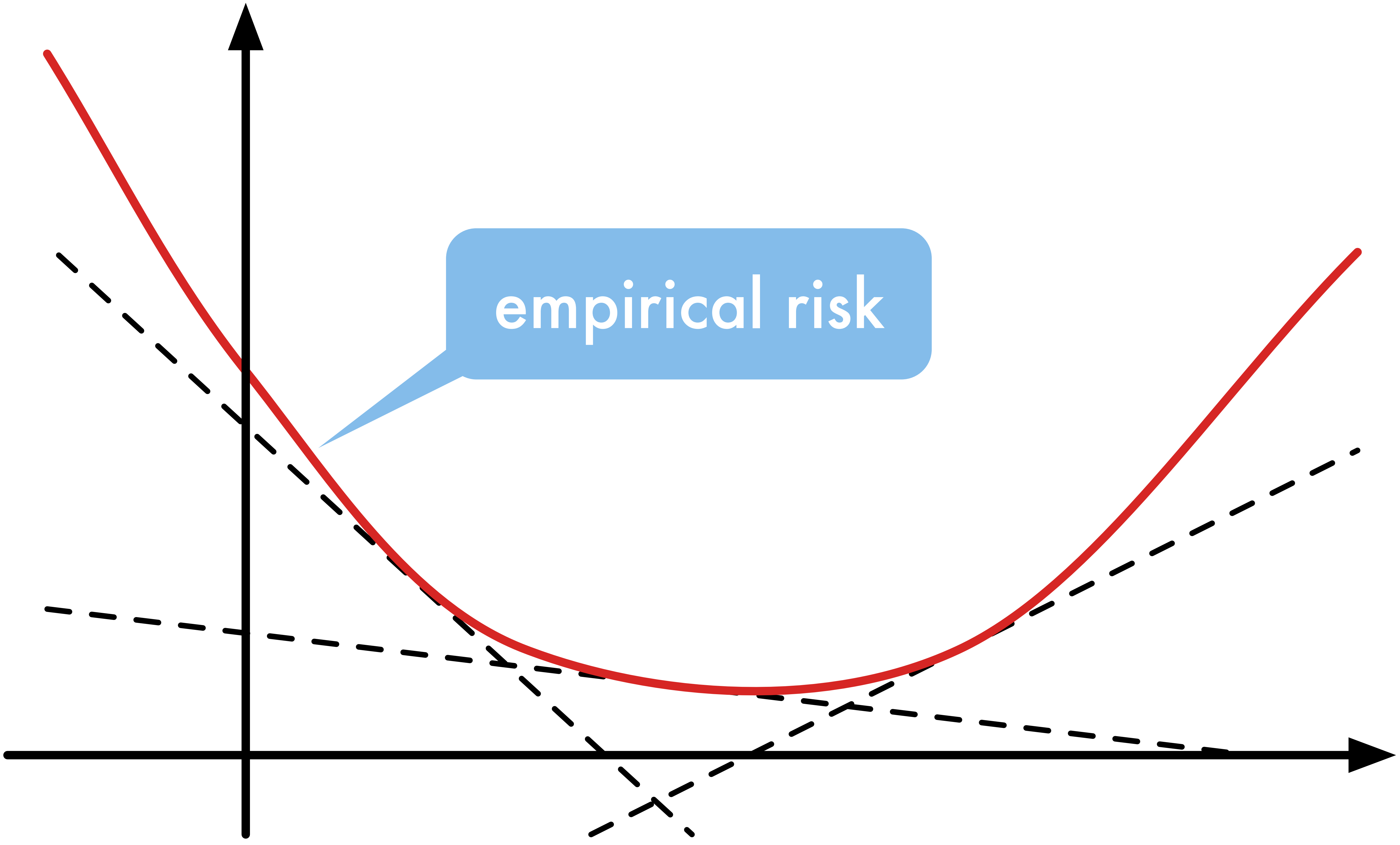
regularizer

Basic Idea

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} l_i(\theta) + \lambda \Omega[\theta]$$

- Loss
  - Convex but expensive to compute
  - Line search just as expensive as new computation
  - Gradient almost free with function value computation
  - Easy to compute in parallel
- Regularizer
  - Convex and cheap to compute and to optimize
- Strategy
  - Compute tangents on loss
  - Provides lower bound on objective
  - Solve dual optimization problem (fewer parameters)

Bundle Method



Lower bound

Regularized Risk Minimization

$$\underset{w}{\text{minimize}}\, R_{\text{emp}}[w] + \lambda\Omega[w]$$

Taylor Approximation for $R_{\text{emp}}[w]$

$$R_{\text{emp}}[w] \geq R_{\text{emp}}[w_t] + \langle w - w_t, \partial_w R_{\text{emp}}[w_t]\rangle = \langle a_t, w\rangle + b_t$$

where $a_t = \partial_w R_{\text{emp}}[w_{t-1}]$ and $b_t = R_{\text{emp}}[w_{t-1}] - \langle a_t, w_{t-1}\rangle$.

Bundle Bound

$$R_{\text{emp}}[w] \geq R_t[w] := \max_{i \leq t} \langle a_i, w\rangle + b_i$$

Regularizer $\Omega[w]$ solves stability problems.

Pseudocode

Initialize $t = 0, w_0 = 0, a_0 = 0, b_0 = 0$
repeat
Find minimizer

$$w_t := \underset{w}{\operatorname{argmin}}\, R_t(w) + \lambda\Omega[w]$$

Compute gradient a_{t+1} and offset b_{t+1}.
Increment $t \leftarrow t + 1$.
until $\epsilon_t \leq \epsilon$

Convergence Monitor $R_{t+1}[w_t] - R_t[w_t]$

Since $R_{t+1}[w_t] = R_{\mathrm{emp}}[w_t]$ (Taylor approximation) we have

$$R_{t+1}[w_t] + \lambda\Omega[w_t] \geq \min_{w} R_{\mathrm{emp}}[w] + \lambda\Omega[w] \geq R_t[w_t] + \lambda\Omega[w_t]$$

Dual Problem

Dual optimization for $\Omega[w] = \frac{1}{2}\|w\|_2^2$ is Quadratic Program regardless of the choice of the empirical risk $R_{\text{emp}}[w]$.

$$\underset{\beta}{\text{minimize}}\ \frac{1}{2\lambda}\beta^\top A A^\top \beta - \beta^\top b$$

$$\text{subject to } \beta_i \geq 0 \text{ and } \|\beta\|_1 = 1$$

The primal coefficient w is given by $w = -\lambda^{-1} A^\top \beta$.

Use Fenchel-Legendre dual of $\Omega[w]$, e.g. $\|\cdot\|_1 \to \|\cdot\|_\infty$.

Can even use simple line search for update (almost as good).

Properties

Parallelization

- Empirical risk sum of many terms: MapReduce
- Gradient sum of many terms, gather from cluster.
- Possible even for multivariate performance scores.
- Data is **local**. Combine data from competing entities.

Solver independent of loss

No need to change solver for new loss.

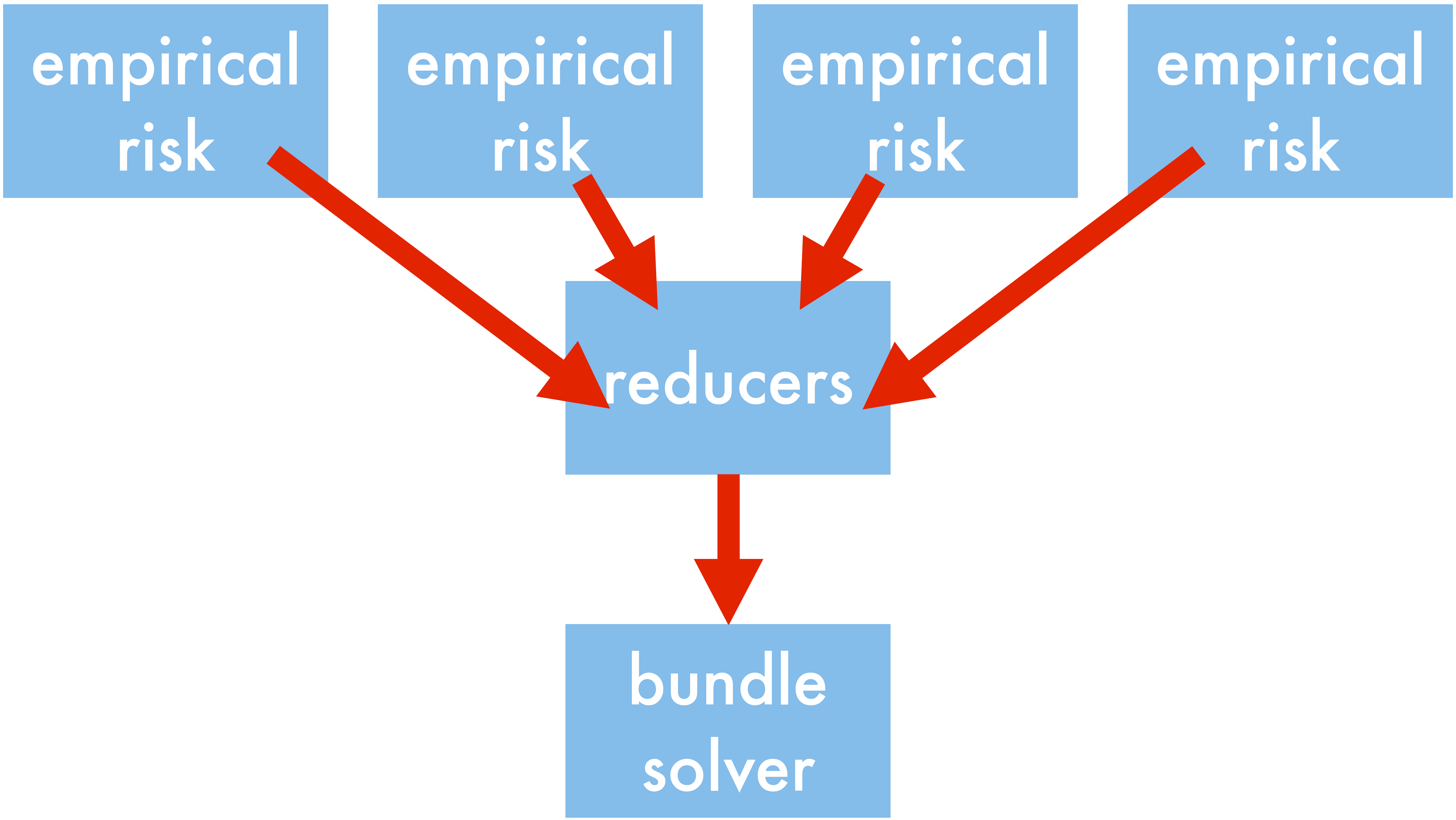
Loss independent of solver/regularizer

Add new regularizer without need to re-implement loss.

Line search variant

- Optimization does not require QP solver at all!
- Update along gradient direction in the dual.
- We only need inner product on gradients!

Implementation



Guarantees

Theorem

The number of iterations to reach ϵ precision is bounded by

$$n \leq \log_2 \frac{\lambda R_{\mathrm{emp}}[0]}{G^2} + \frac{8G^2}{\lambda\epsilon} - 4$$

steps. If the Hessian of $R_{\mathrm{emp}}[w]$ is bounded, convergence to any $\epsilon \leq \lambda/2$ takes at most the following number of steps:

$$n \leq \log_2 \frac{\lambda R_{\mathrm{emp}}[0]}{4G^2} + \frac{4}{\lambda} \max\left[0, 1 - 8G^2 H^*/\lambda\right] - \frac{4H^*}{\lambda} \log 2\epsilon$$

Advantages

- Linear convergence for smooth loss
- For non-smooth loss almost as good in practice (as long as smooth on a course scale).
- Does not require primal line search.

Proof idea

Duality Argument

- Dual of $R_i[w] + \lambda\Omega[w]$ lower bounds minimum of regularized risk $R_{\mathrm{emp}}[w] + \lambda\Omega[w]$.
- $R_{i+1}[w_i] + \lambda\Omega[w_i]$ is upper bound.
- **Show that the gap $\gamma_i := R_{i+1}[w_i] - R_i[w_i]$ vanishes.**

Dual Improvement

- Give lower bound on increase in dual problem in terms of γ_i and the subgradient $\partial_w \left[R_{\mathrm{emp}}[w] + \lambda\Omega[w]\right]$.
- For unbounded Hessian we have $\delta\gamma = O(\gamma^2)$.
- For bounded Hessian we have $\delta\gamma = O(\gamma)$.

Convergence

- Solve difference equation in γ_t to get desired result.

4.3 Online Methods

Stochastic gradient descent

- Empirical risk as expectation

$$\frac{1}{m}\sum_{i=1}^{m} l\left(y_i - \langle \phi(x_i), \theta \rangle\right) = \mathbf{E}_{i \sim \{1,..m\}}\left[l\left(y_i - \langle \phi(x_i), \theta \rangle\right)\right]$$

- Stochastic gradient descent (pick random x,y)

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \partial_\theta \left(y_t, \langle \phi(x_t), \theta_t \rangle\right)$$

- Often we require that parameters are restricted to some convex set X, hence we project on it

$$\theta_{t+1} \leftarrow \pi_x \left[\theta_t - \eta_t \partial_\theta \left(y_t, \langle \phi(x_t), \theta_t \rangle\right)\right]$$

$$\text{here } \pi_X(\theta) = \operatorname*{argmin}_{x \in X} \|x - \theta\|$$

Convergence in Expectation

initial loss

$$\mathbf{E}_{\bar\theta}\left[l(\bar\theta)\right] - l^* \leq \frac{R^2 + L^2 \sum_{t=0}^{T-1} \eta_t^2}{2\sum_{t=0}^{T-1} \eta_t} \text{ where}$$

$$l(\theta) = \mathbf{E}_{(x,y)}\left[l(y, \langle \phi(x), \theta\rangle)\right] \text{ and } l^* = \inf_{\theta \in X} l(\theta) \text{ and } \bar\theta = \frac{\sum_{t=0}^{T-1} \theta_t \eta_t}{\sum_{t=0}^{T-1} \eta_t}$$

expected loss

parameter average

- Proof
  Show that parameters converge to minimum

$$\theta^* \in \operatorname*{argmin}_{\theta \in X} l(\theta) \text{ and set } r_t := \|\theta^* - \theta_t\|$$

from Nesterov and Vial

Proof

$$
\begin{aligned}
r_{t+1}^2 &= \left\|\pi_X[\theta_t - \eta_t g_t] - \theta^*\right\|^2 \\
&\leq \left\|\theta_t - \eta_t g_t - \theta^*\right\|^2 \\
&= r_t^2 + \eta_t^2 \left\|g_t\right\|^2 - 2\eta_t \left\langle \theta_t - \theta^*, g_t \right\rangle \\
\text{hence } \mathbf{E}\left[r_{t+1}^2 - r_t^2\right] &\leq \eta_t^2 L^2 + 2\eta_t \left[l^* - \mathbf{E}[l(\theta_t)]\right] \\
&\leq \eta_t^2 L^2 + 2\eta_t \left[l^* - \mathbf{E}[l(\bar{\theta})]\right]
\end{aligned}
$$

by convexity

- Summing over inequality for t proves claim
- This yields randomized algorithm for minimizing objective functions (try log times and pick the best / or average median trick)

Rates

- Guarantee

$$\mathbf{E}_{\bar{\theta}}\left[l(\bar{\theta})\right] - l^* \leq \frac{R^2 + L^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{t=0}^{T-1} \eta_t}$$

- If we know R, L, T pick constant learning rate

$$\eta = \frac{R}{L\sqrt{T}} \text{ and hence } \mathbf{E}_{\bar{\theta}}[l(\bar{\theta})] - l^* \leq \frac{R[1 + 1/T]L}{2\sqrt{T}} < \frac{LR}{\sqrt{T}}$$

- If we don't know T pick $\eta_t = O(t^{-\frac{1}{2}})$
  This costs us an additional log term

$$\mathbf{E}_{\bar{\theta}}[l(\bar{\theta})] - l^* = O\left(\frac{\log T}{\sqrt{T}}\right)$$

Strong Convexity

$$l_i(\theta') \geq l_i(\theta) + \langle \partial_\theta l_i(\theta), \theta' - \theta \rangle + \frac{1}{2}\lambda \left\| \theta - \theta' \right\|^2$$

- **Use this to bound the expected deviation**

$$\begin{aligned} r_{t+1}^2 &\leq r_t^2 + \eta_t^2 \left\| g_t \right\|^2 - 2\eta_t \langle \theta_t - \theta^*, g_t \rangle \\ &\leq r_t^2 + \eta_t^2 L^2 - 2\eta_t \left[l_t(\theta_t) - l_t(\theta^*) \right] - 2\lambda \eta_t r_k^2 \end{aligned}$$

$$\text{hence } \mathbf{E}[r_{t+1}^2] \leq (1 - \lambda h_t)\mathbf{E}[r_t^2] - 2\eta_t \left[\mathbf{E}\left[l(\theta_t) \right] - l^* \right]$$

- **Exponentially decaying averaging**

$$\bar{\theta} = \frac{1-\sigma}{1-\sigma^T} \sum_{t=0}^{T-1} \sigma^{T-1-t} \theta_t$$

and plugging this into the discrepancy yields

$$l(\bar{\theta}) - l^* \leq \frac{2L^2}{\lambda T} \log \left[1 + \frac{\lambda R T^{\frac{1}{2}}}{2L} \right] \text{ for } \eta = \frac{2}{\lambda T} \log \left[1 + \frac{\lambda R T^{\frac{1}{2}}}{2L} \right]$$

More variants

- Adversarial guarantees

$$\theta_{t+1} \leftarrow \pi_x \left[\theta_t - \eta_t \partial_\theta \left(y_t, \langle \phi(x_t), \theta_t \rangle\right)\right]$$

  has low regret (average instantaneous cost) for arbitrary orders (useful for game theory)

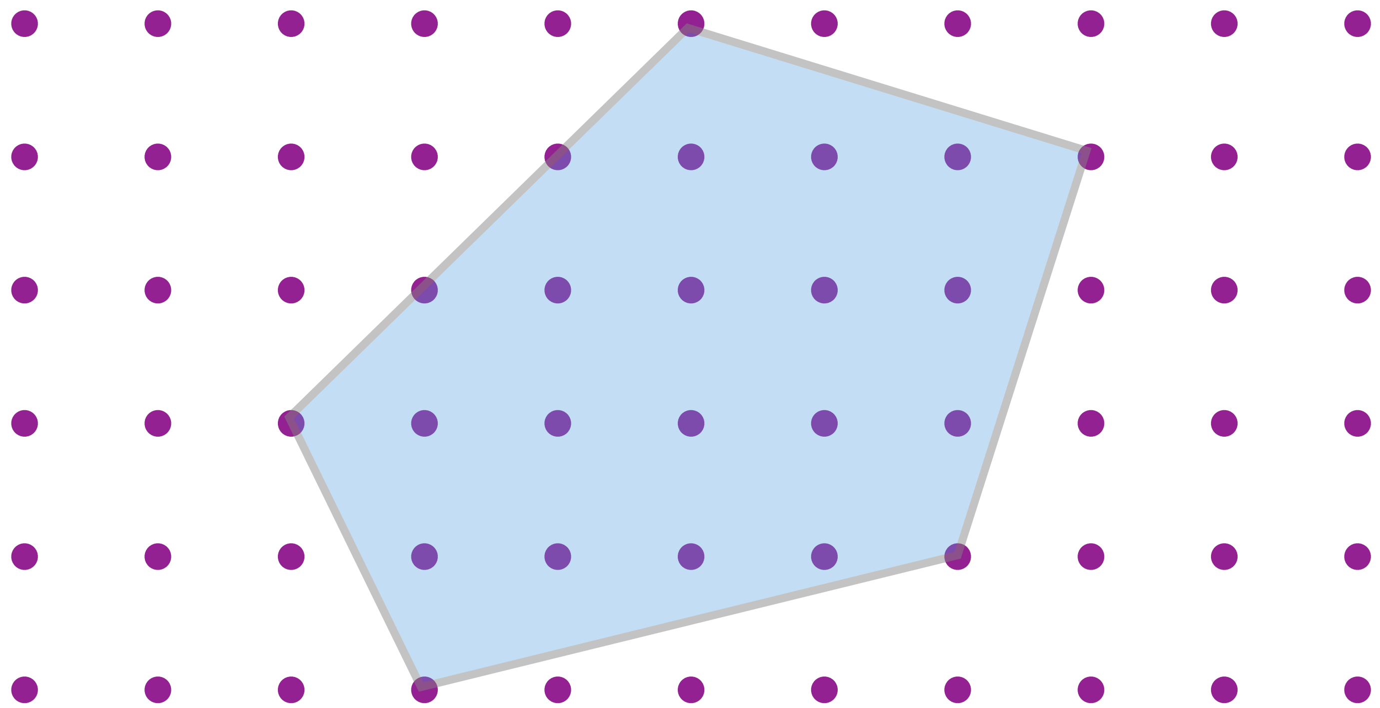
  - Ratliff, Bagnell, Zinkevich
    $O(t^{-\frac{1}{2}})$ learning rate
  - Shalev-Shwartz, Srebro, Singer (Pegasos)
    $O(t^{-1})$ learning rate (but need constants)
  - Bartlett, Rakhlin, Hazan
    (add strong convexity penalty)

4.4 Discrete Problems

Integer programming relaxations

- Optimization problem

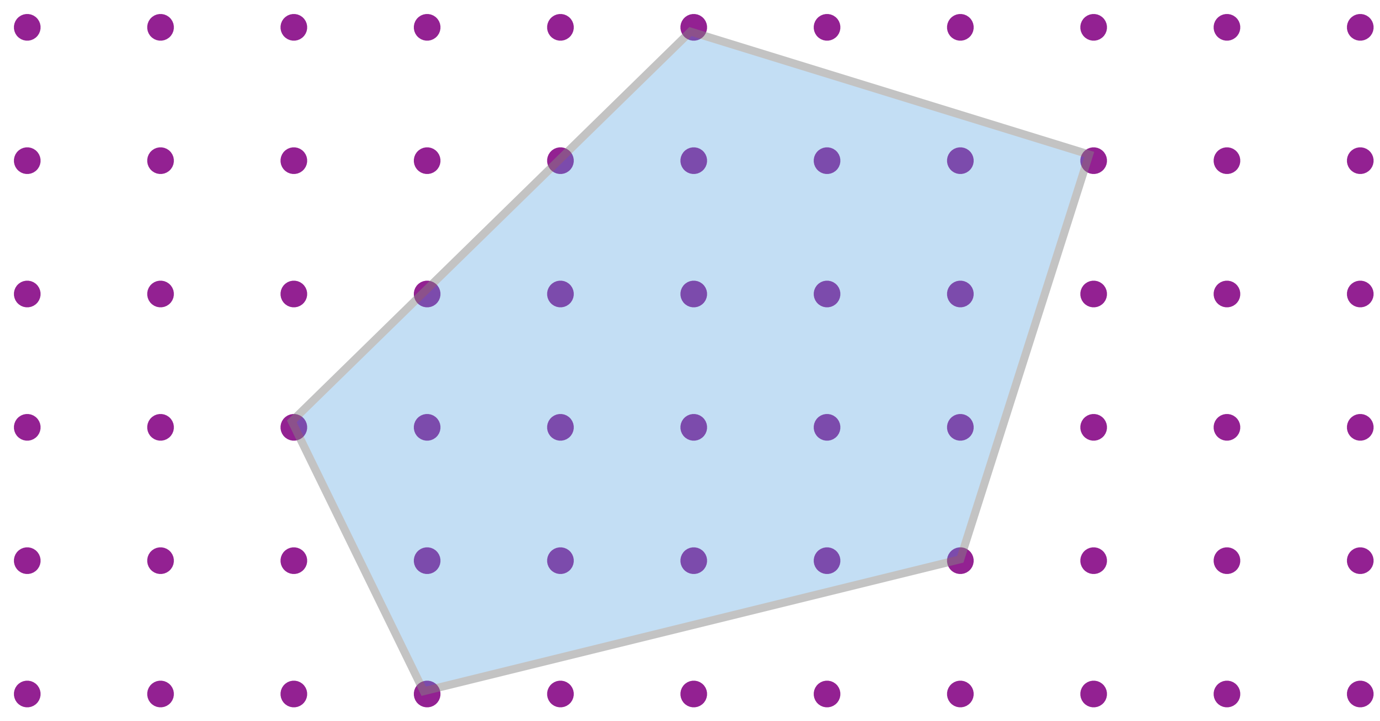
$$\underset{x}{\text{minimize}}\ c^\top x \text{ subject to } Ax \leq b \text{ and } x \in \mathbb{Z}^n$$



- Relax to linear program if vertices are integral since LP has vertex solution

Integer programming relaxations

- Totally unimodular constraint matrix A
  - Inverse of each submatrix must be integral
  - RHS of constraints must be integral
  - Many useful sufficient conditions for TU.



Example - Hungarian Marriage

- Optimization Problem
  - n Hungarian men
  - n Hungarian women
  - Compatibility c_{ij} between them
- Find optimal matching

$$\underset{\pi}{\text{maximize}} \sum_{ij} \pi_{ij} C_{ij}$$

$$\text{subject to } \pi_{ij} \in \{0, 1\} \text{ and } \sum_i \pi_{ij} = 1 \text{ and } \sum_j \pi_{ij} = 1$$

- All vertices of the constraint matrix are integral

Randomization

- Maximum finding
  - Very large set of instances
  - Find approximate maximum

- Draw a random set of n terms
- Take maximum over subset
  (59 for 95% with 95% confidence)

$$\Pr\left\{F[\max_i x_i] < \epsilon\right\} = (1-\epsilon)^n = \delta$$

$$\text{hence } n = \frac{\log \delta}{\log(1-\epsilon)} \leq \frac{-\log \delta}{\epsilon}$$

Randomization

- Find good solution
  - Show that expected value is well behaved
  - Show that tails are bounded
  - Sufficiently large random draw must contain at least one good element (e.g. CM sketch)
- Find good majority
  - Show that majority satisfies condition
  - Bound probability of minority being overrepresented (e.g. Mean-Median theorem)
- Much more in these books
  - Raghavan & Motwani (Randomized Algorithms)
  - Alon & Spencer (Probabilistic Method)

Submodular maximization

- Submodular function
  - Defined on sets
  - Diminishing returns property

$$f(A \cup C) - f(A) \geq f(B \cup C) - f(B) \text{ for } A \subseteq B$$

- Example
  For web search results we might have individually

 >  >  >  >  > 

But if we can show only 4 we should probably pick

 +  +  + 

Submodular maximization

- Optimization problem

$$\max_{X \in \mathcal{X}} f(X) \text{ subject to } |X| \leq k$$

  Often NP hard even to find tight approximation

- Greedy optimization procedure
  - Start with empty set X
  - Find x such that $f(X \cup \{x\})$ is maximized
  - Add x to the set and repeat until |X|=k
  - Guarante of (1 - 1/e) optimality

Further reading

- Nesterov and Vial (expected convergence)
  http://dl.acm.org/citation.cfm?id=1377347
- Bartlett, Hazan, Rakhlin (strong convexity SGD)
  http://books.nips.cc/papers/files/nips20/NIPS2007_0699.pdf
- TAO (toolkit for advanced optimization)
  http://www.mcs.anl.gov/research/projects/tao/
- Ratliff, Bagnell, Zinkevich
  http://martin.zinkevich.org/publications/ratliff_nathan_2007_3.pdf
- Shalev-Shwartz, Srebro, Singer (Pegasos paper)
  http://dl.acm.org/citation.cfm?id=1273598
- Langford, Smola, Zinkevich (slow learners are fast)
  http://arxiv.org/abs/0911.0491
- Hogwild (Recht, Wright, Re)
  http://pages.cs.wisc.edu/~brecht/papers/hogwildTR.pdf