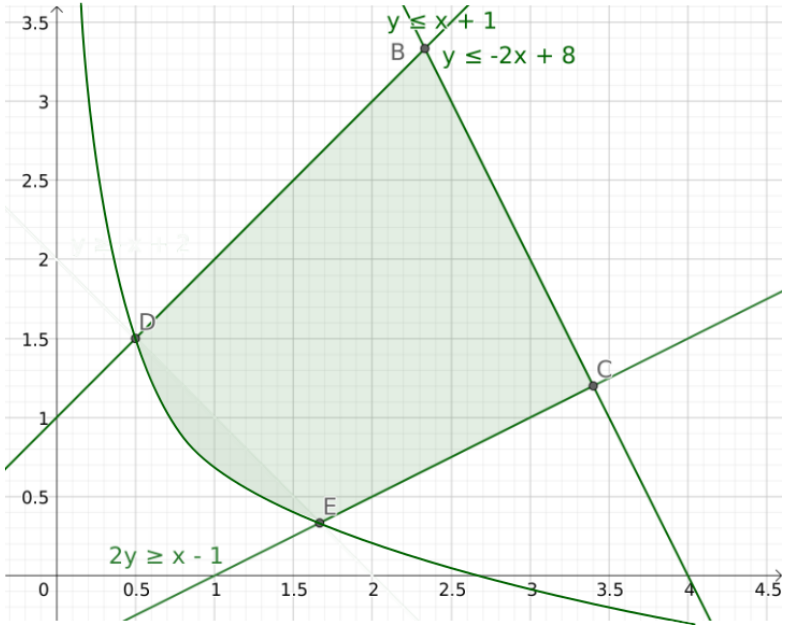
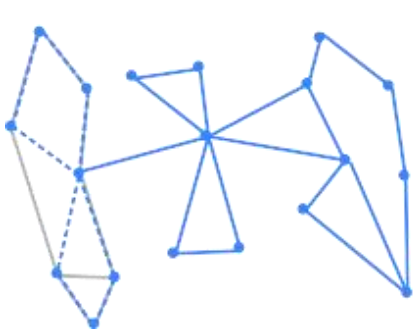
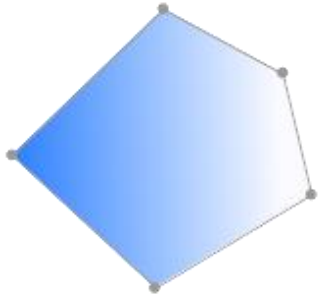
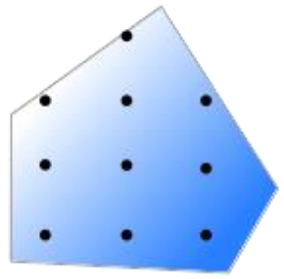
Convex Optimisation 3

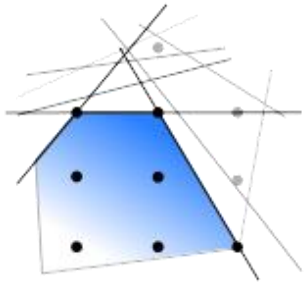
COMP4691 / 8691

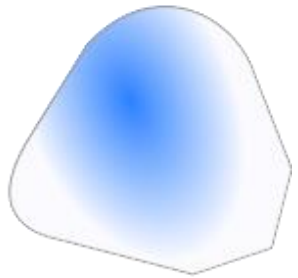






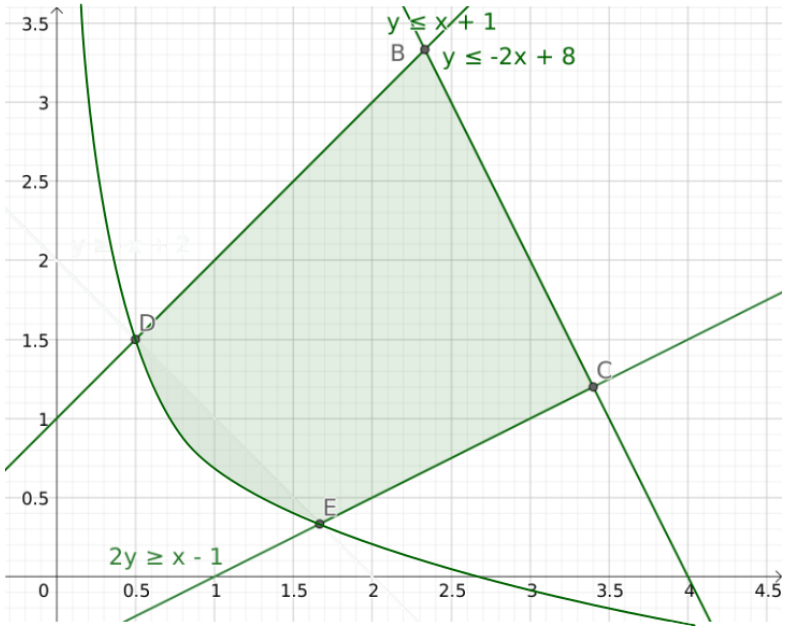






Convex Optimisation Outline

- Convexity
- Unconstrained Optimisation
- Constrained Optimisation
- **Lagrangian Duality**
  - Recap
  - Relationship with LP Duality
  - Dual Gradient Ascent
- KKT Conditions
- Interior Point Method



In some cases we will look at more general non-convex problems, as some of the theory applies there also.

Lagrangian Dual Problem

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

Dual variables, one for each constraint again

Also known as **Lagrange** or **KKT multipliers**

Lagrangian Function:

$$\mathcal{L}(x, \mu, \lambda) = f(x) + \mu^\intercal g(x) + \lambda^\intercal h(x) \qquad \mu \geq 0$$

Lagrangian Dual Function:

$$d(\mu, \lambda) = \inf_x \mathcal{L}(x, \mu, \lambda)$$

The dual function is **always** concave!

like a min, see next slide

Dual Problem:

$$\max_{\mu \geq 0, \lambda} d(\mu, \lambda) = \max_{\mu \geq 0, \lambda} \inf_x \mathcal{L}(x, \mu, \lambda)$$

Find tightest lower bound

Lagrangian Dual Problem

Weak duality holds: $\max_{\mu \geq 0, \lambda} d(\mu, \lambda) \leq \min_{x} f(x)$ s.t. $g(x) \leq 0$, $h(x) = 0$

Functions don't need to be differentiable, convex, or even continuous.

In order to get **strong duality** we need some more conditions.

One example is that strong duality holds for a convex optimisation problem when **Slater's Condition** holds.

For convex problems, Slater's Condition is **sufficient**, but **not necessary** for **strong duality** to hold (there are other conditions).

For **most convex** optimisation problems you come across will have **strong duality**.

Lagrangian Dual Problem: LP

I've said the LP dual problem is actually a Lagrangian dual, but it looks quite different… let's demonstrate the connection

$$\min_x \ c^\intercal x \quad \text{s.t. } Ax \leq b \qquad f(x) = c^\intercal x \qquad g(x) = Ax - b$$

$$\mathcal{L}(x, \mu) = c^\intercal x + \mu^\intercal (Ax - b)$$

$$\max_{\mu \geq 0} \inf_x \mathcal{L}(x, \mu)$$

We can utilise necessary condition (unconstrained) for inner problem:

$$\begin{aligned} & \nabla_x \mathcal{L}(x, \mu) = 0 \\ \implies & \nabla_x c^\intercal x + \nabla_x \mu^\intercal (Ax - b) = 0 \\ \implies & c^\intercal + \mu^\intercal A = 0 \\ \implies & A^\intercal \mu = -c \end{aligned}$$

Eliminating c from Lagrangian function:

$$\begin{aligned} \mathcal{L}(x, \mu) &= c^\intercal x + \mu^\intercal (Ax - b) \\ &= -(A^\intercal \mu)^\intercal x + \mu^\intercal (Ax - b) \\ &= -\mu^\intercal b \end{aligned}$$

This is also sufficient as LP is convex.

Lagrangian Dual Problem: LP

$$A^\intercal \mu = -c \implies \inf_x \mathcal{L}(x, \mu) = -\mu^\intercal b$$
$$A^\intercal \mu \neq -c \implies \inf_x \mathcal{L}(x, \mu) \text{ unbounded below}$$

Which of these cases do we need to consider?

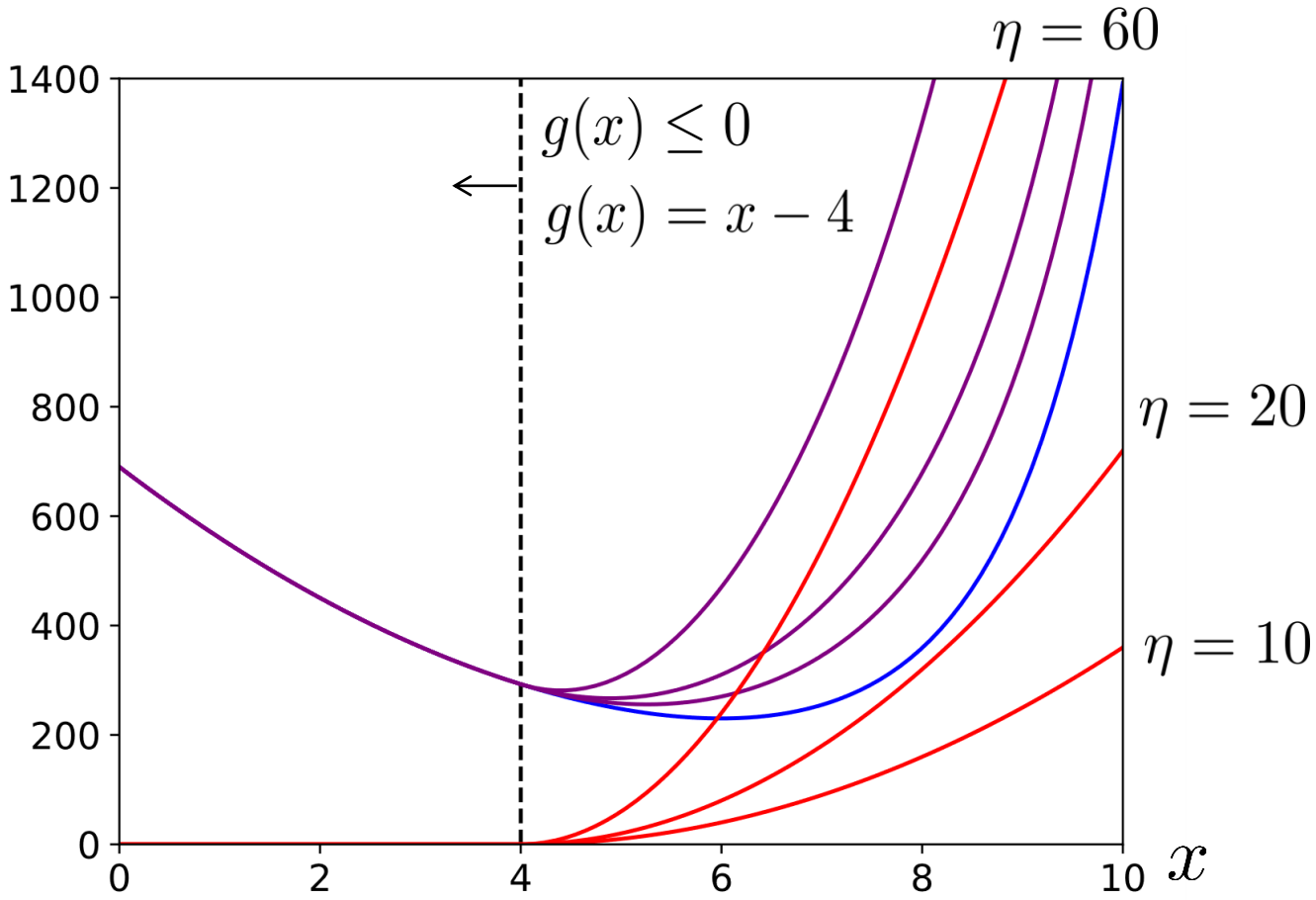
$$\max_{\mu \geq 0} \inf_x \mathcal{L}(x, \mu) = \max_{\mu \geq 0} \; -\mu^\intercal b$$
$$\text{s.t. } A^\intercal \mu = -c$$
$$= \max_y \; b^\intercal y$$
$$\text{s.t. } A^\intercal y = c$$
$$y \leq 0$$

Variable change: μ to -y

The Lagrangian Dual problem simplifies to solving another LP!

Recap: Penalty Method

$$x^{(k)} = \arg\min_x \; f(x) + \eta_k \| \max(0, g(x)) \|_2^2 + \eta_k ||h(x)||_2^2$$



The penalty needs to approach infinity in order to satisfy the constraint

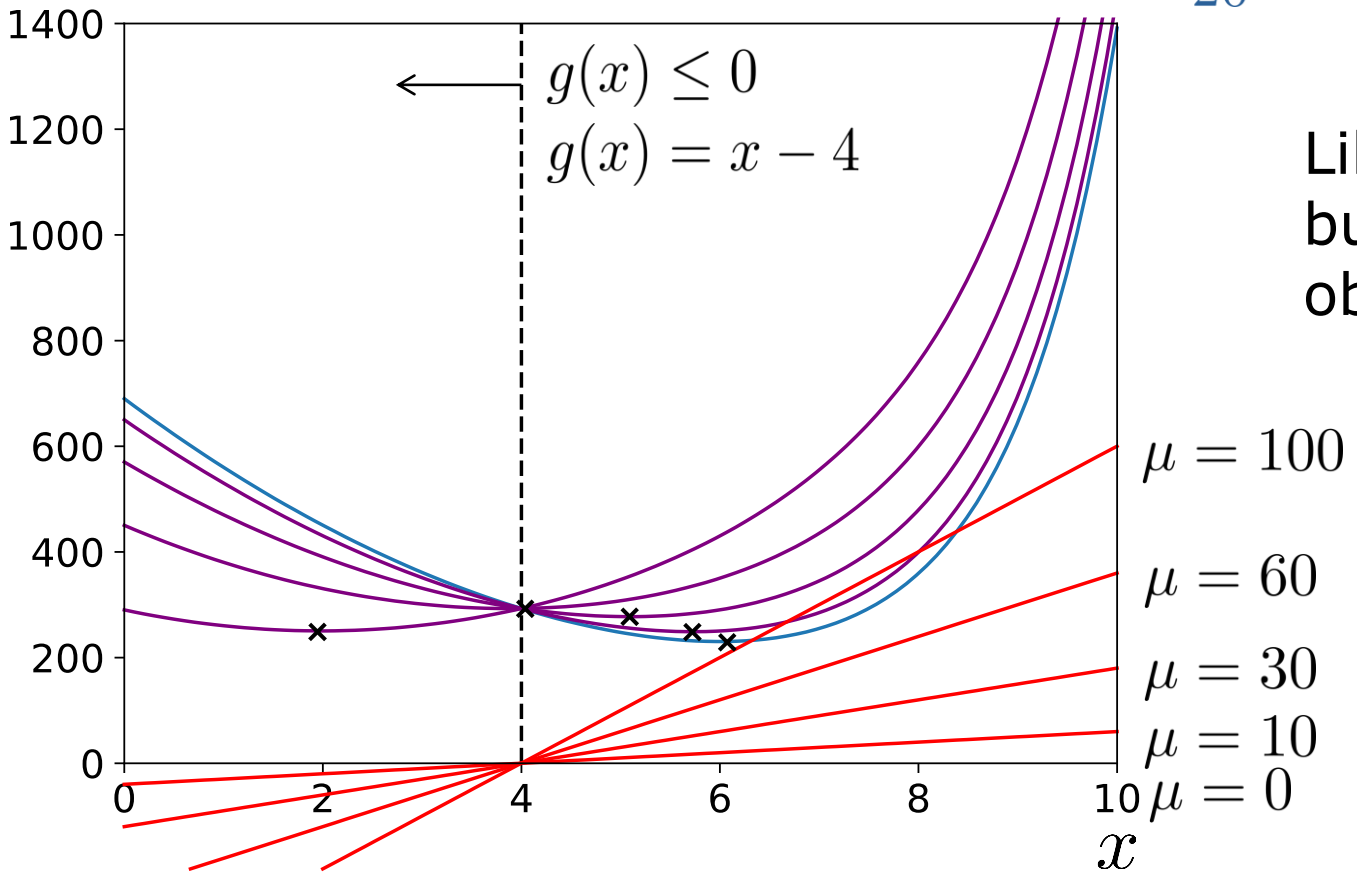
$$f(x) = 10(x-7)^2 + \frac{1}{20}e^x + 200$$

Details omitted… we can do something better…

Recap: Lagrangian Relaxation

The Lagrangian relaxation of g:

$$\mathcal{L}(x, \mu) = 10(x-7)^2 + \frac{1}{20}e^x + 200 + \mu(x-4)$$



Like the penalty method penalty, but linear, and also impacts objective within feasible region.

This hints at an alternative algorithm for solving constrained problems.

Dual Gradient Ascent

$$\min_x \; f(x)$$
$$\text{s.t. } g(x) \leq 0$$

$$\max_{\mu \geq 0} d(\mu) = \max_{\mu \geq 0} \inf_x \mathcal{L}(x, \mu)$$

$$\mathcal{L}(x, \mu) = f(x) + \mu^\intercal g(x)$$

Iteratively solve the dual problem. Because strong duality holds for (most) convex problems, we can get an answer to the original problem.

$$x_{k+1} := \arg\min_x \mathcal{L}(x, \mu_k)$$

(we need existence of a min now)

$$\mu_{k+1} := \mu_k + \alpha_k \nabla_\mu d(\mu_k) = \mu_k + \alpha_k \nabla_\mu \mathcal{L}(x_{k+1}, \mu_k)$$
$$= \mu_k + \alpha_k g(x_{k+1})$$

Note: dual function not direction

A gradient **ascent** step

Also need to ensure that the μ dual variables remain positive.

Dual Gradient Ascent

$$x_{k+1} := \arg\min_{x} \mathcal{L}(x, \mu_k)$$
$$\mu_{k+1} := \mu_k + \alpha_k g(x_{k+1})$$

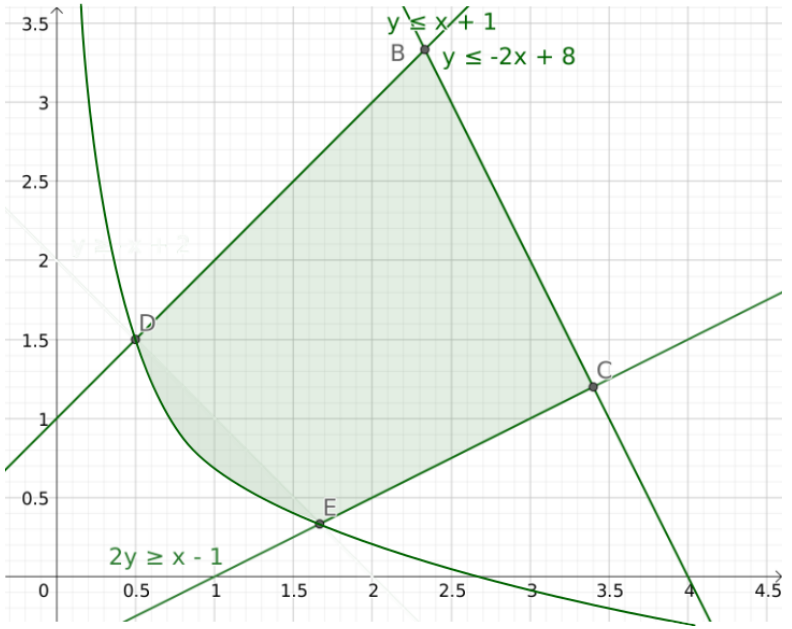
This only works when:

- the dual function is differentiable
- the problem is strictly convex (Lagrangian can be unbounded below for particular dual values)

Other first-order methods, including subgradient-based methods and the the *method of multipliers* can be used instead for convex problems if these conditions don't hold.

Convex Optimisation Outline

- Convexity
- Unconstrained Optimisation
- Constrained Optimisation
- Lagrangian Duality
- **KKT Conditions**
- Interior Point Method



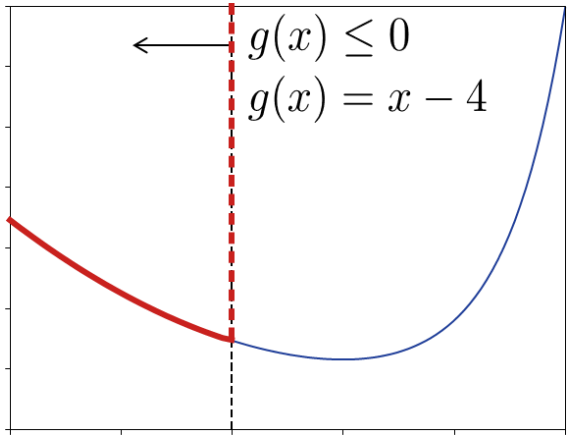
In some cases we will look at more general non-convex problems, as some of the theory applies there also.

Constrained Optimisation

For unconstrained optimisation we relied on the necessary condition that a local minimum will be at a point with zero gradient.

$$\nabla f(x^*) = 0 \quad H_f(x^*) \succeq 0$$

This is no longer necessary for constraint problems!



KKT conditions are the analogous necessary conditions for optimality in a constrained context.

KKT Conditions

Karush-Kuhn-Tucker (KKT) Conditions are:

First-order necessary conditions for a **local optimal** in a constrained problem, so long as some **regularity conditions** are met.

Regularity conditions: the problem satisfies certain rules that restrict the form of the functions and constraints used.

For **convex** optimisation problems **Slater's condition** is a sufficient (not necessary) regularity condition; however, there are more general ones.

1951 **Tucker + Kuhn**: "birth" of nonlinear programming with KKT conditions
1939 **Karush**: had these conditions in his masters thesis

Sufficient Conditions

There also exist **second-order sufficient conditions** for a **local optimal** that are analogous to those in an unconstrained setting. We will not cover them here.

When we have a **convex problem**, whose functions are continuously differentiable, the **KKT conditions alone become sufficient** (also for a broader class of functions called *invex* functions).

KKT Conditions

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

$$\mathcal{L}(x, \mu, \lambda) = f(x) + \mu^\intercal g(x) + \lambda^\intercal h(x)$$

Assume all **functions are differentiable** (again subderivative forms exist).

For a constrained problem, that satisfies some regularity conditions, the first order necessary conditions (KKT conditions) are:

not going to fit on this slide...

KKT Conditions

For x^* to be a local minimum, there exist some KKT multipliers μ λ where:

$$\nabla_x \mathcal{L}(x^*, \mu, \lambda) = 0 \qquad \text{(stationarity)}$$

$$g(x^*) \leq 0 \qquad h(x^*) = 0 \qquad \text{(primal feasibility)}$$

$$\mu \geq 0 \qquad \text{(dual feasibility)}$$

$$\mu_j g_j(x^*) = 0 \qquad \text{(complementary slackness)}$$

$$\nabla f(x^*) + J_g(x^*)^\intercal \mu + J_h(x^*)^\intercal \lambda = 0$$

$$\nabla f(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) = 0$$

KKT Interpretation

We can consider the dual variables as the force at which a constraint pushes back preventing us from getting a better solution.

Stationarity: A stationary point in x for the Lagrangian. The "force" due to the objective is exactly balanced by constraints preventing a better solution.

$$\nabla f(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) = 0$$

Primal feasibility: We are in the feasible region.

Dual feasibility: An inequality can only block in one direction. $\mu \geq 0$

Complementary Slackness: Constraint can only push back if we are touching it. $\mu_j g_j(x^*) = 0$

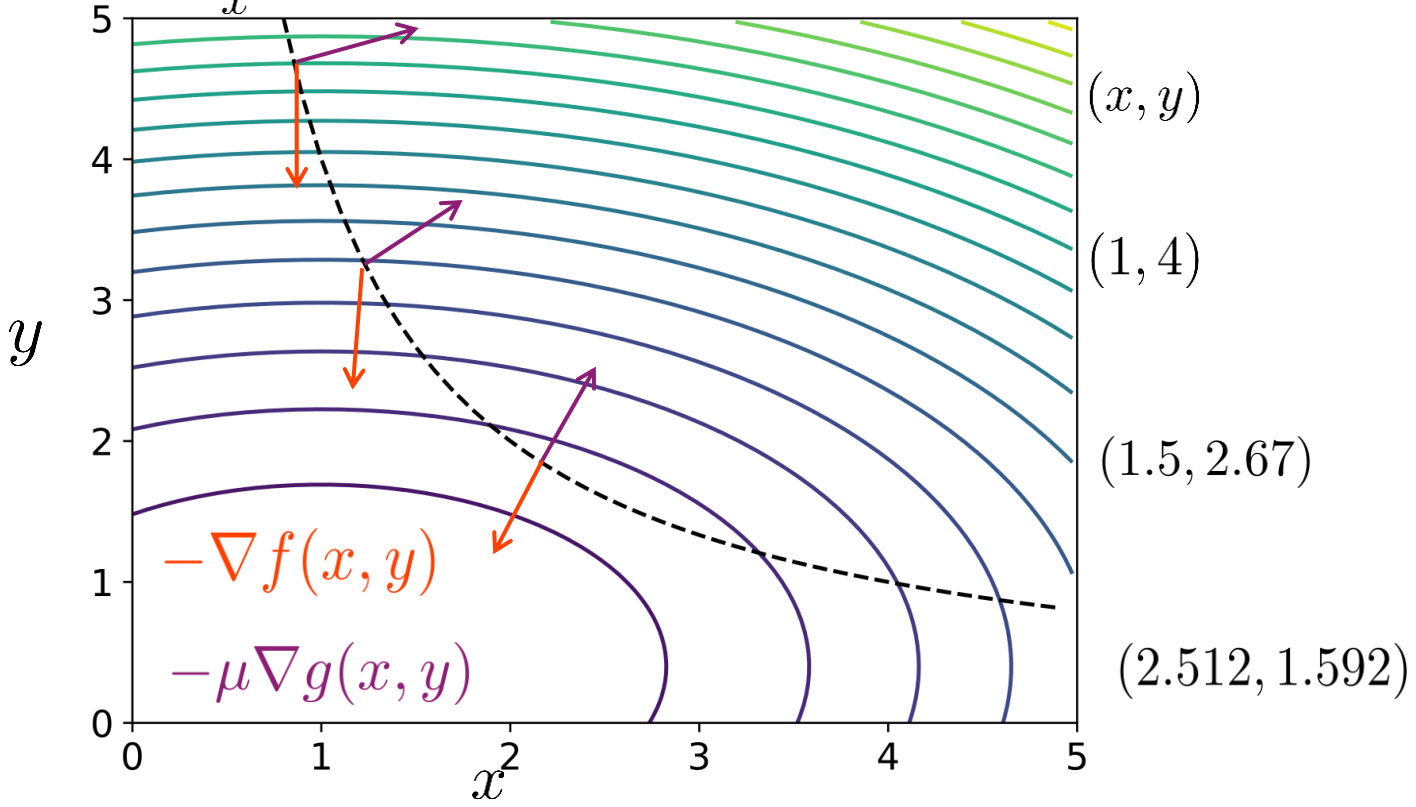
KKT Stationarity

$$g(x,y) \le 0$$

$$g(x,y) = \frac{4}{x} - y$$

$$y \ge \frac{4}{x}$$

$$f(x,y) = 3(x-1)^2 + 6(y-0.4)^2$$



$$\nabla_{x\ldots y}\mathcal{L}(x,y,\mu) = \nabla f(x,y) + \mu\nabla g(x,y)$$

$$\nabla f(x,y) = \begin{bmatrix} 6(x-1) \\ 12(y-0.4) \end{bmatrix} \quad \nabla g(x,y) = \begin{bmatrix} -\frac{4}{x^2} \\ -1 \end{bmatrix}$$

$$\nabla f(x,y) + \mu\nabla g(x,y) = 0$$

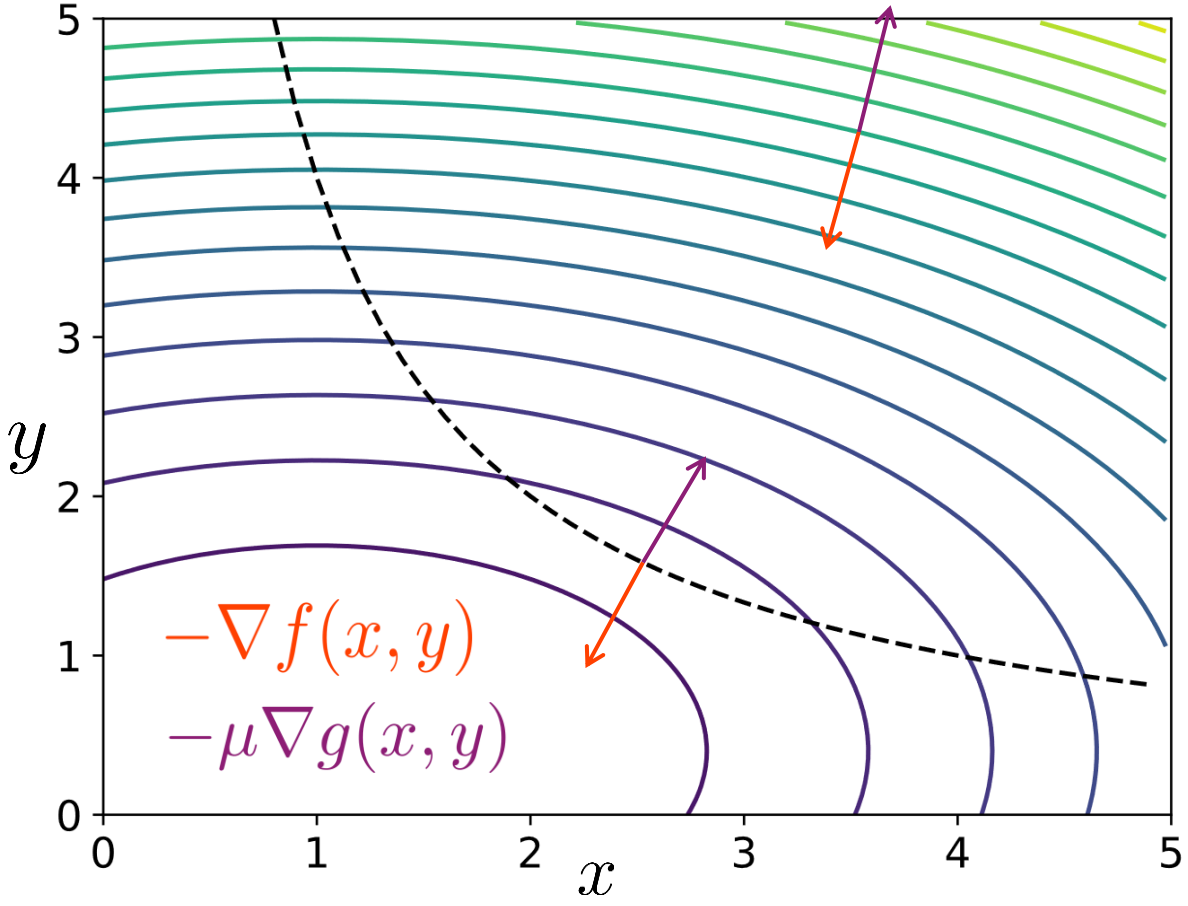
$$0 - 4\mu = 0 \implies \mu = 0$$
$$43.2 - \mu = 0 \implies \mu = 43.2$$

$$3 - 1.78\mu = 0 \implies \mu = 1.69$$
$$27.2 - \mu = 0 \implies \mu = 27.2$$

$$9.07 - 0.634\mu = 0 \implies \mu = 14.3$$
$$14.3 - \mu = 0 \implies \mu = 14.3$$

KKT Stationarity

$$g(x,y) \leq 0 \qquad g(x,y) = \frac{4}{x} - y \quad \longleftarrow \quad y \geq \frac{4}{x} \qquad f(x,y) = 3(x-1)^2 + 6(y-0.4)^2$$



$$\nabla f(x,y) = \begin{bmatrix} 6(x-1) \\ 12(y-0.4) \end{bmatrix} \qquad \nabla g(x,y) = \begin{bmatrix} -\frac{4}{x^2} \\ -1 \end{bmatrix}$$

Complementary slackness:

$$\mu g(x,y) = 0$$

$$\mu\left(\frac{4}{x} - y\right) = 0$$

Prevents solutions such as...

Solving KKT

For problems where KKT is sufficient, can we just **directly solve for these conditions** rather than the two stage dual ascent approach?

Stationary points of the Lagrangian function are saddle points rather than minima or maxima. Some algorithms like gradient descent / ascent will struggle to find them.

Newton's method can find saddle points; however, we still typically don't directly try to solve the KKT conditions. **Interior point method** next lecture.

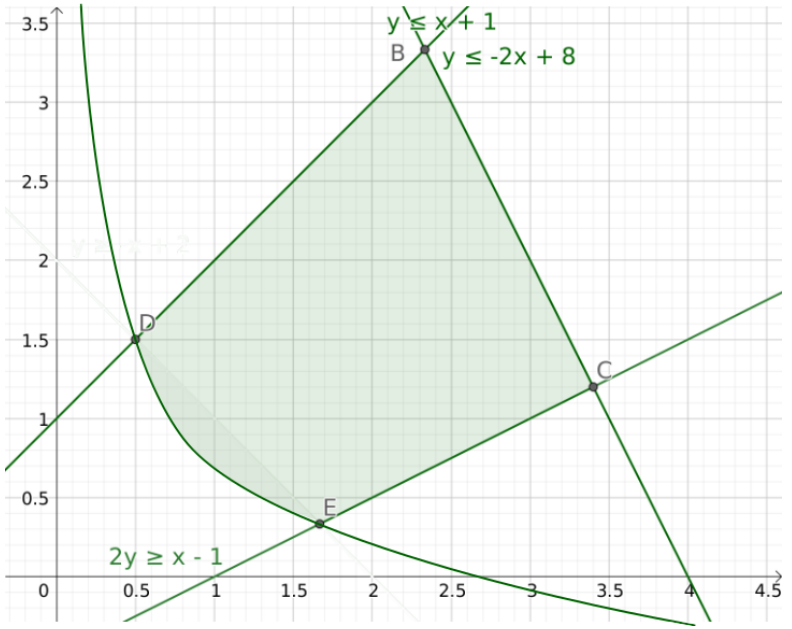
KKT for Nonconvex

KKT necessary for all local minima, whether convex or not, when the problem meets regularity conditions.

If we could enumerate all of them we could pick the global optimal, but just finding one is a challenge, and it might not be known how many exist for a problem or where to look.

Convex Optimisation Outline

- Convexity
- Unconstrained Optimisation
- Constrained Optimisation
- Lagrangian Duality
- KKT Conditions
- **Interior Point Method**



In some cases we will look at more general non-convex problems, as some of the theory applies there also.

Interior Point Method

Barrier functions introduced to objective that approach infinity at inequality constraint boundary.

Progressively move towards a solution that satisfies the KKT conditions, rather than attempting to directly solve them.

We will outline the algorithm (with some simplifications) that lies behind **Ipopt**. This is a **primal-dual** interior point method. More details:

Wachter and Biegler 2006 *On the Implementation of an Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming*

Generate unconstraint problem and over time we ensure that the problem remains feasible

Barrier Functions

Approach infinity as the variables approach the boundary of a constraint.

$$\min_x \ f(x)$$
$$\text{s.t. } g(x) \leq 0$$

Some barrier functions:

$$-\log(-g_i(x))$$
$$\frac{1}{-g_i(x)}$$

Smooths out the sharp transition of a constraint.

Barrier problem for a given ν : → Barrier parameter

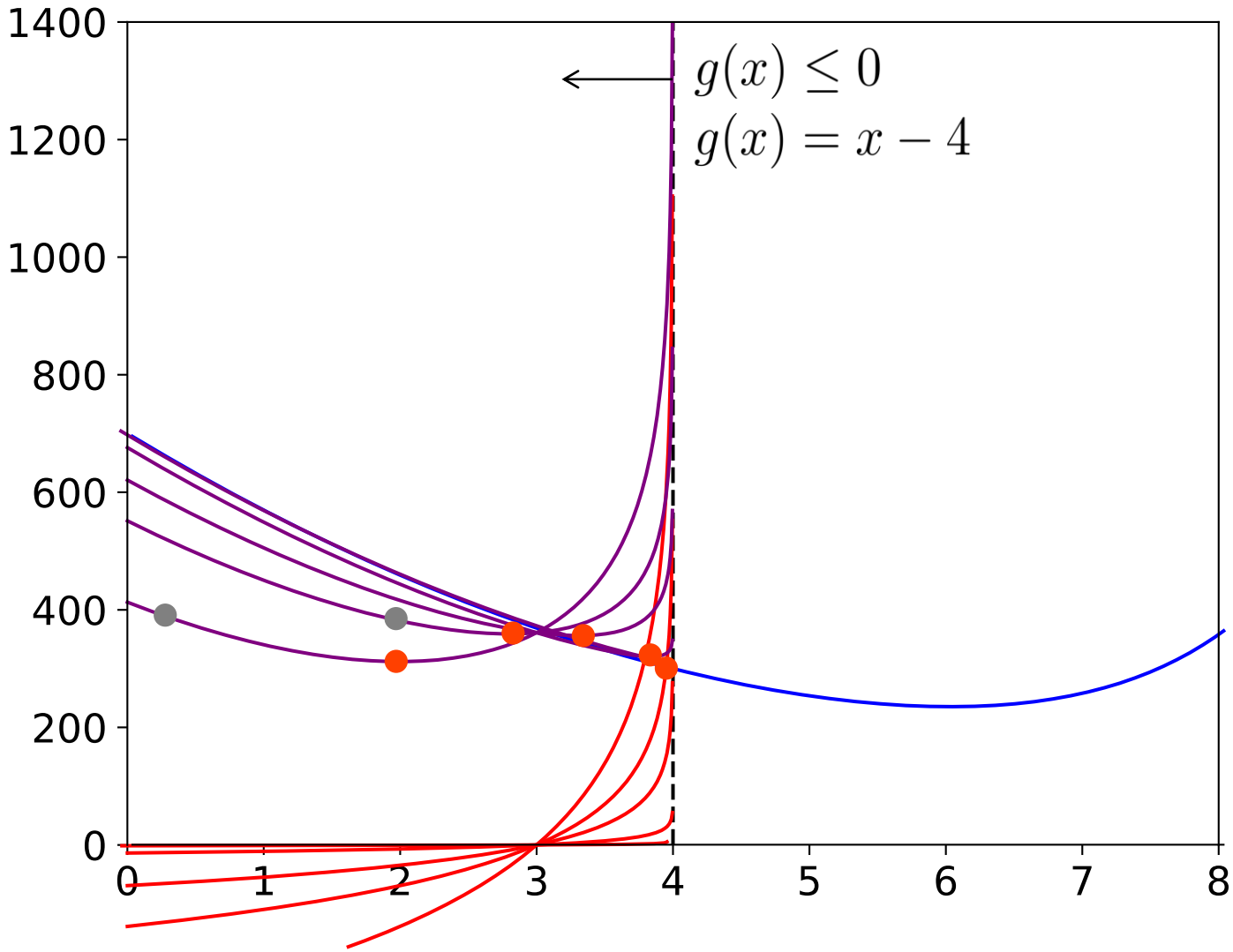
$$\min_x \ \phi(x, \nu) := f(x) - \nu \sum_i^m \log(-g_i(x))$$

Objective +barrier function

The **barrier method** (interior point method), solves a sequence of barrier problems with ν converging to zero.

Barrier Method Visualised

$$f(x) = 10(x-7)^2 + \frac{1}{20}e^x + 200$$



$$\phi(x, \upsilon) = f(x) - \upsilon \log(4 - x)$$

$\nu = 200$

$\nu = 100$

$\nu = 50$

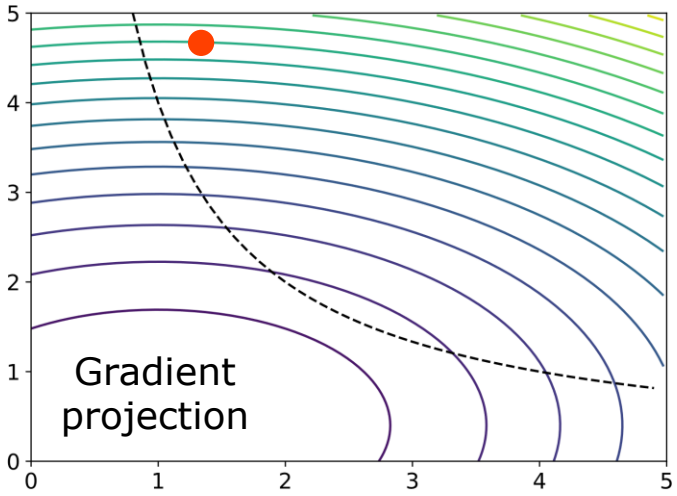
$\nu = 10$

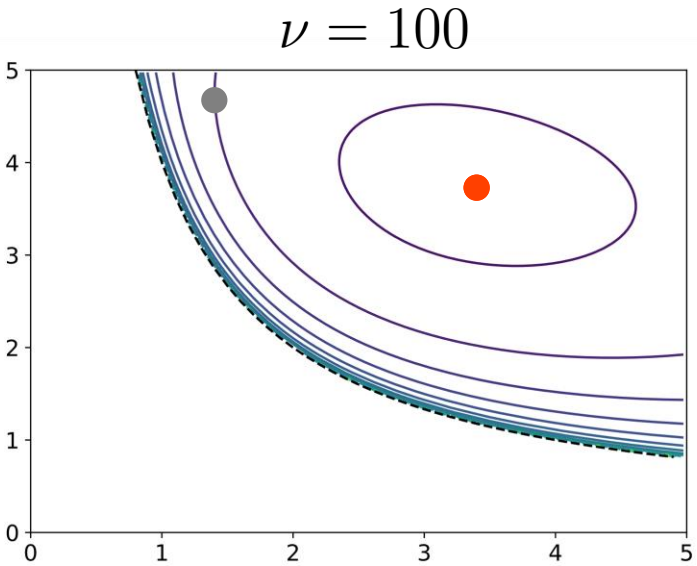
$\nu = 1$

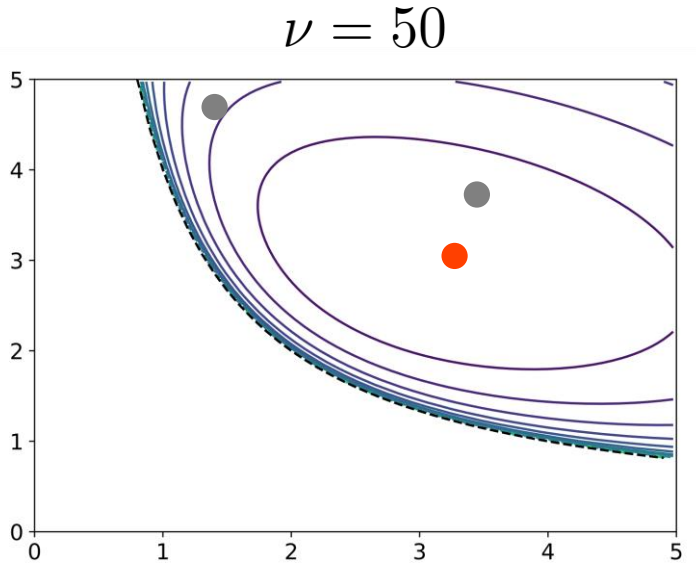
As we get to the edge of our constraint the barrier leads to infinity

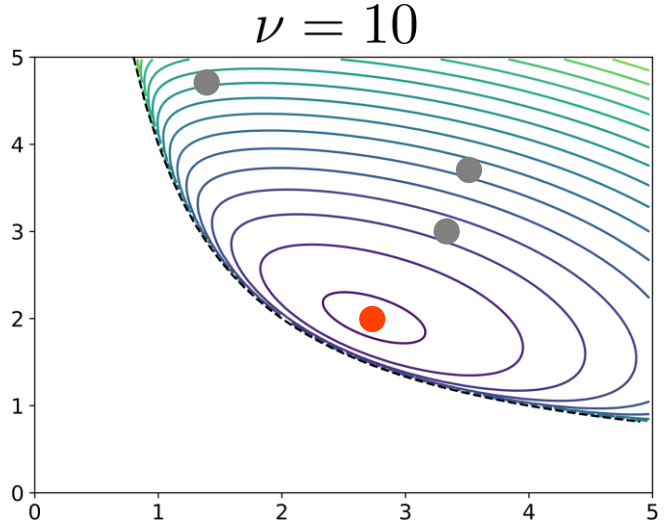
Nice smooth transition to the optimal.

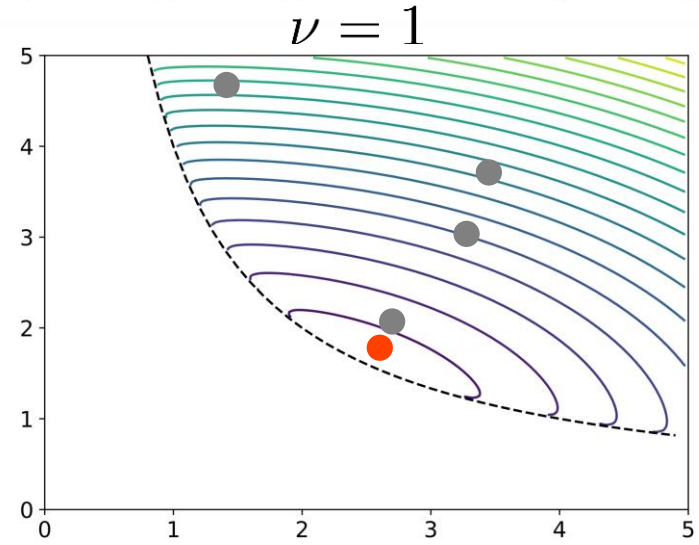
Barrier Method Visualised

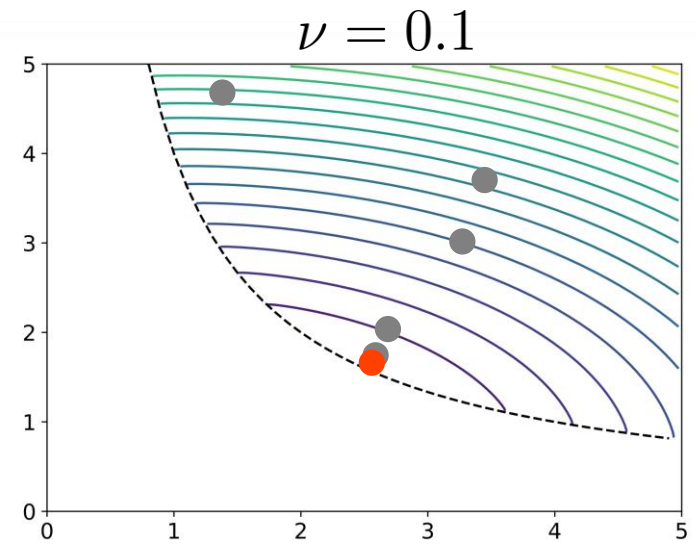












Primal-Dual IPM / BM

We can gradually solve the **primal** and **dual** variables for the problem, warmstarting the barrier problem with both for each update of ν

We are going to solve problem in the following form:

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & h(x) = 0 \\ & x \geq 0 \end{aligned}$$

Barrier functions applied to just inequality constraints:

$$\begin{aligned} \min_x \ & \phi(x, \nu) := f(x) - \nu \sum_j^n \ln(x_j) \\ \text{s.t. } & h(x) = 0 \end{aligned}$$

Complementary Slackness ($\mu_j g_j(x^*) = 0$) is not need for problems in this form

The barrier problem is an equality **constrained problem**, that we solve by applying Newton's Method to its **KKT conditions**.

Newton's Method (Review)

For finding roots of f:

$$\nabla f(x_k)^\intercal (x_{k+1} - x_k) = -f(x_k) \qquad \text{if f scalar valued}$$

$$J_f(x_k)(x_{k+1} - x_k) = -f(x_k) \qquad \text{if f vector valued}$$

To find **stationary points instead of zeros**, we can replace f with its gradient (assuming it is twice differentiable):

$$J_{\nabla f}(x_k)(x_{k+1} - x_k) = -\nabla f(x_k)$$

$$\implies H_f(x_k)(x_{k+1} - x_k) = -\nabla f(x_k)$$

Remember Jacobian of gradient is Hessian.

When adapted to finding zeros of the gradient, it becomes a **second-order** method.

Solving the Barrier Problem

$$\min_x \; f(x) - \nu \sum_j^n \ln(x_j)$$

$$\text{s.t. } h(x) = 0$$

KKT conditions: $\nabla_x \mathcal{L}(x, \lambda) = 0$

$h(x) = 0$

greatly simplified when no inequalities

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla f(x) - \nu \sum_j^n \nabla \ln(x_j) + \sum_i^m \lambda_i \nabla h_i(x)$$

$$= \nabla f(x) - \nu x^{-1} + \sum_i^m \lambda_i \nabla h_i(x)$$

where we define the "inverse" of a vector notation as: $x_j^{-1} = \frac{1}{x_j}$

Solving the Barrier Problem

$$s(x,\lambda) := \nabla f(x) - \nu x^{-1} + \sum_{i}^{m} \lambda_i \nabla h_i(x) = 0$$
$$h(x) = 0$$

We employ Newton's method to iteratively solve this system of nonlinear equations.

$$J_u(x_k,\lambda_k)\begin{bmatrix} x_{k+1} - x_k \\ \lambda_{k+1} - \lambda_k \end{bmatrix} = -u(x_k,\lambda_k) \qquad u(x_k,\lambda_k) := \begin{bmatrix} s(x_k,\lambda_k) \\ h(x_k) \end{bmatrix}$$

$$(n+m)\times(n+m)$$
$$J_u(x_k,\lambda_k) = \begin{bmatrix} J_{s,x}(x_k,\lambda_k) & J_{s,\lambda}(x_k,\lambda_k) \\ J_{h,x}(x_k) & J_{h,\lambda}(x_k) \end{bmatrix} = \begin{bmatrix} W_k & A_k^\intercal \\ A_k & 0 \end{bmatrix}$$

$$W_k := H_{\mathcal{L},x}(x_k,\lambda_k) \qquad A_k := J_{h,x}(x_k)$$

Stepping

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} - x_k \\ \lambda_{k+1} - \lambda_k \end{bmatrix} = - \begin{bmatrix} s(x_k, \lambda_k) \\ h(x_k) \end{bmatrix}$$

A linear system of equations (once we plug the iterate values in)

Generally sparse

Solving for this (the search direction)

Line search on primal and dual variables.

Once barrier problem is solved to within a certain tolerance, the barrier parameter ν is reduced, and the new barrier problem is solved, warmstarting from the above primal and dual values.

Link to Original KKT

$$s(x,\lambda) := \nabla f(x) - \nu x^{-1} + \sum_i^m \lambda_i \nabla h_i(x) = 0$$

$$\mu = \nu x^{-1}$$

$$\nabla f(x) - \mu + \sum_i^m \lambda_i \nabla h_i(x) = 0$$

Gradient of Lagrangian

$$\mu_j x_j = \nu$$

We can rewrite the KKT conditions for the barrier problem as:

$$\nabla f(x) - \mu + \sum_i^m \lambda_i \nabla h_i(x) = 0$$

$$h(x) = 0$$

$$\mu_j x_j = \nu$$

As ν approaches zero, and keeping $x \geq 0 \quad \mu \geq 0$, these become the KKT conditions for the original problem!

$$g(x) := -x \leq 0$$

The barrier method gradually enforces the nasty complementary slackness term.

First vs Second Order Methods

The interior point method explained here is a **second-order** method.

These methods typically need **fewer iterations** to converge, but more work per iteration (e.g., calculating the Hessian and solving a large linear system).

For **very large** (e.g., big data) applications, **first-order** methods are often necessary, as the Hessian can become too big to represent in memory and factorise. They also present more opportunity to efficiently distribute the computation.

That said, for small to medium problems (up to hundreds of thousands of variables) and complicated constraints, interior point methods are often superior.