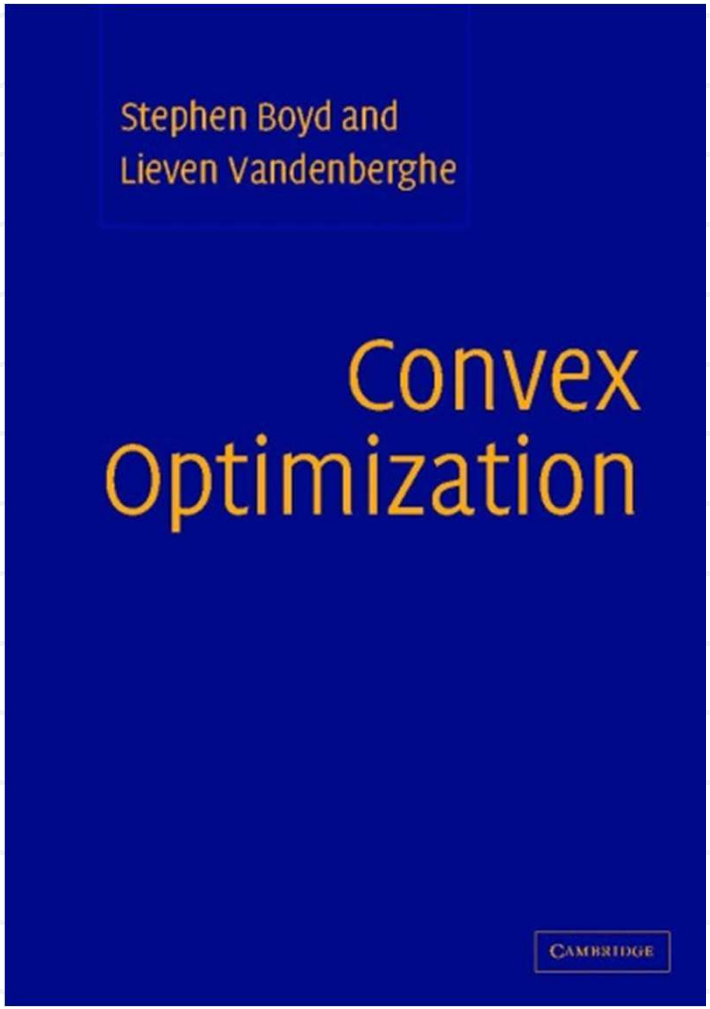
Lecture 15: Subgradient optimization. KKT conditions.

Reading



Section 5.5

Recap (again)

The Lagrangian:

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

The dual function:

$$g(\lambda,\nu) = \inf_x L(x,\lambda,\nu)$$

Let x to be any feasible primal set of variables, and (λ, ν) be any feasible dual set of variables (i.e. $\lambda \geq 0$):

$$g(\lambda,\nu) \leq L(x,\lambda,\nu) \leq f_0(x)$$

In particular, this holds for the optimal values:

$$g(\lambda^*,\nu^*) \leq L(x^*,\lambda^*,\nu^*) \leq f_0(x^*)$$

Partial dualization

$$\min f_0(x) \quad x \in D$$
$$\text{s.t. } f_i(x) \le 0, \; i = 1 \dots m$$
$$h_i(x) = 0, \; i = 1 \dots p$$

$$x \in D'$$
$$f_i'(x) \le 0, \; i = 1 \dots m'$$
$$h_i'(x) = 0, \; i = 1 \dots p'$$

The (partial) Lagrangian:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

Observation: suppose $x \in D$ and $\lambda \ge 0$. Then

$$L(x, \lambda, \nu) \le f_0(x)$$

Consequence: $\forall \lambda \ge 0, \nu$

$$\min_{x \in D'} L(x, \lambda, \nu) \le \min_{x \in D \cap D'} L(x, \lambda, \nu) \le \min_{x \in D \cap D'} f_0(x)$$

The Lagrange dual (partial dualization)

The (partial) dual function:

$$g(\lambda, \nu) = \min_{x \in D'} L(x, \lambda, \nu)$$

Theorem (*weak duality*):

$$\max_{\lambda \geq 0, \nu} g(\lambda, \nu) \leq \min_{f_i(x) \leq 0,\ f_i'(x) \leq 0,\ h_i(x) = 0,\ h_i'(x) = 0} f_0(x)$$

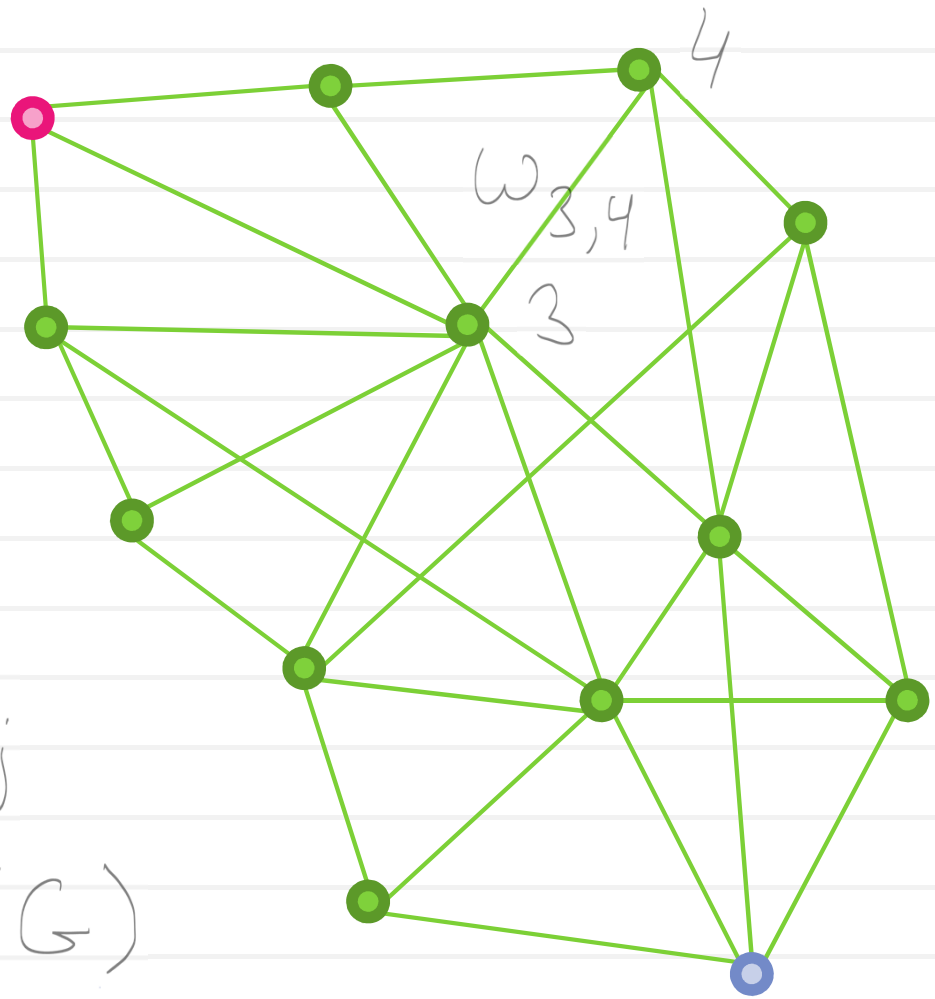
We have a freedom to pick which constraints to dualize and which to leave as they are

Constrained shortest path

Shortest path task

- Dijkstra
- Bellman-Ford
- Breadth-first search

.....



$$\min_{x} \sum w_{ij} \cdot x_{ij}$$

$$s.t. \quad x \in Path(G)$$

$$x_{ij} \in \{0;1\}$$

$$\sum x_{ij} = K$$

Constrained shortest path

$$\min_x \sum \omega_{ij} \cdot x_{ij}$$

$$\text{s.t.} \quad x \in \text{Path}(G)$$

$$\sum x_{ij} = k$$

$$g(\lambda) = \min_{x \in \text{Path}(G)} \left(\sum \omega_{ij} \cdot x_{ij} + \lambda \left(k - \sum x_{ij} \right) \right)$$

$$= \lambda k + \min_{x \in \text{Path}(G)} \left(\sum (\omega_{ij} - \lambda) \cdot x_{ij} \right)$$

Dual can be evaluated using shortest path!

Constrained shortest path

$$g(\lambda) = \lambda K + \min_{x \in Path(G)} \left(\sum (\omega_{ij} - \lambda) \cdot x_{ij} \right)$$

$$\frac{dg}{d\lambda} = \left(K - \sum \hat{x}_{ij} \right)$$

$$\lambda_{t+1} = \lambda_t + \delta_t \cdot \frac{dg}{d\lambda}$$

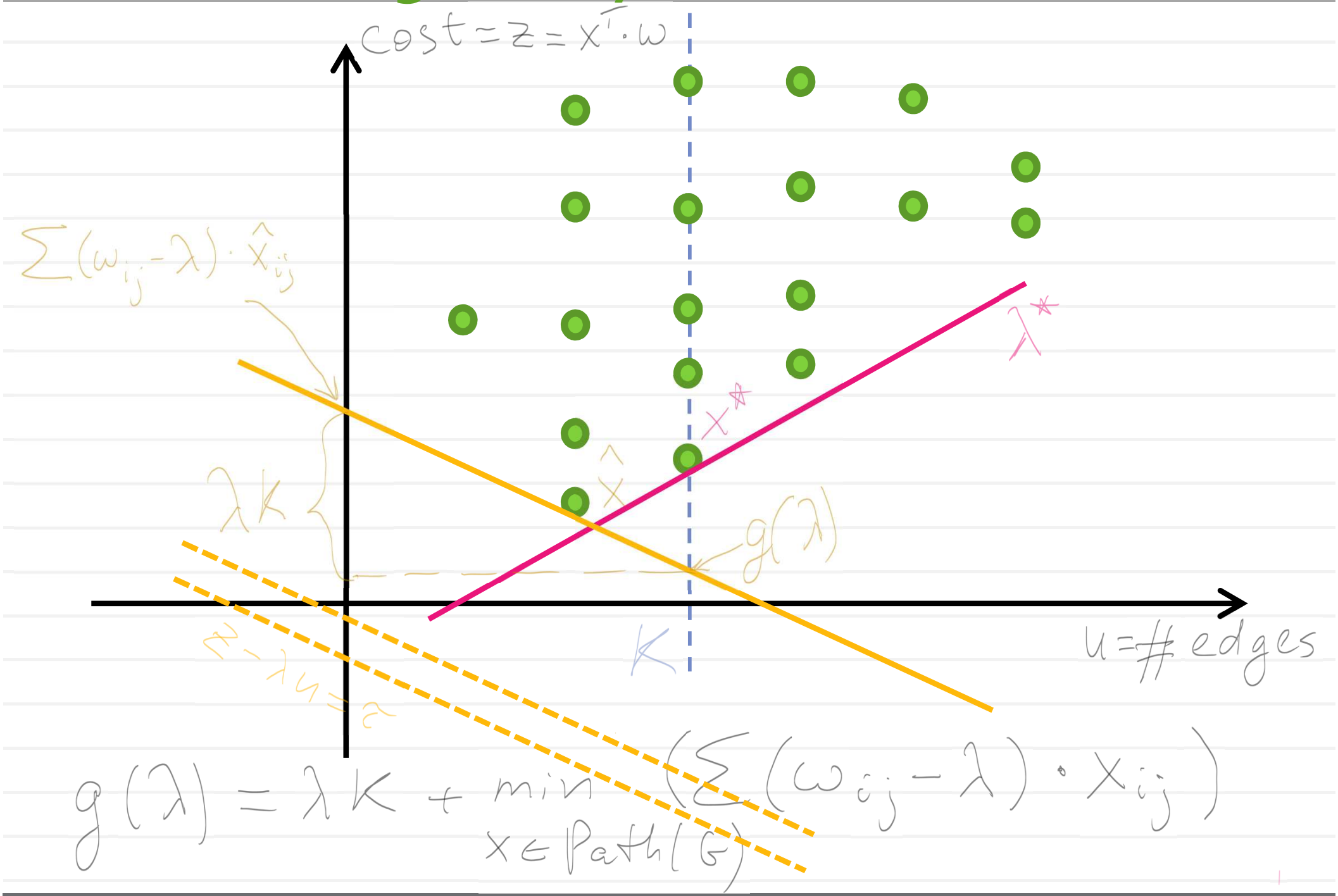
$$\delta_t \to 0$$

$$\sum \delta_t = +\infty$$

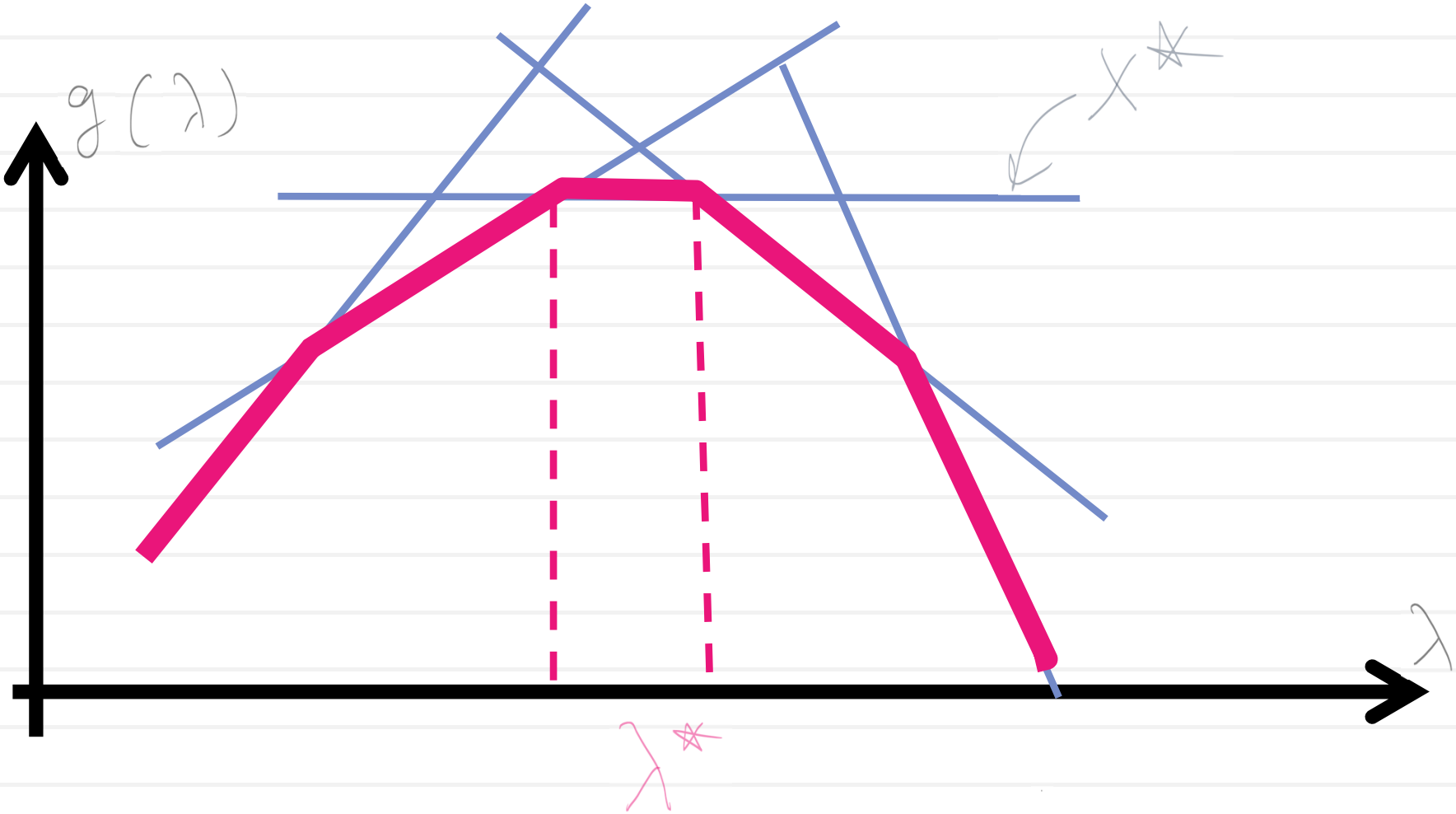
Subgradient ascent step in this case:

- If the optimal path is longer than K, decrease λ
- If the optimal path is shorter than K, increase λ

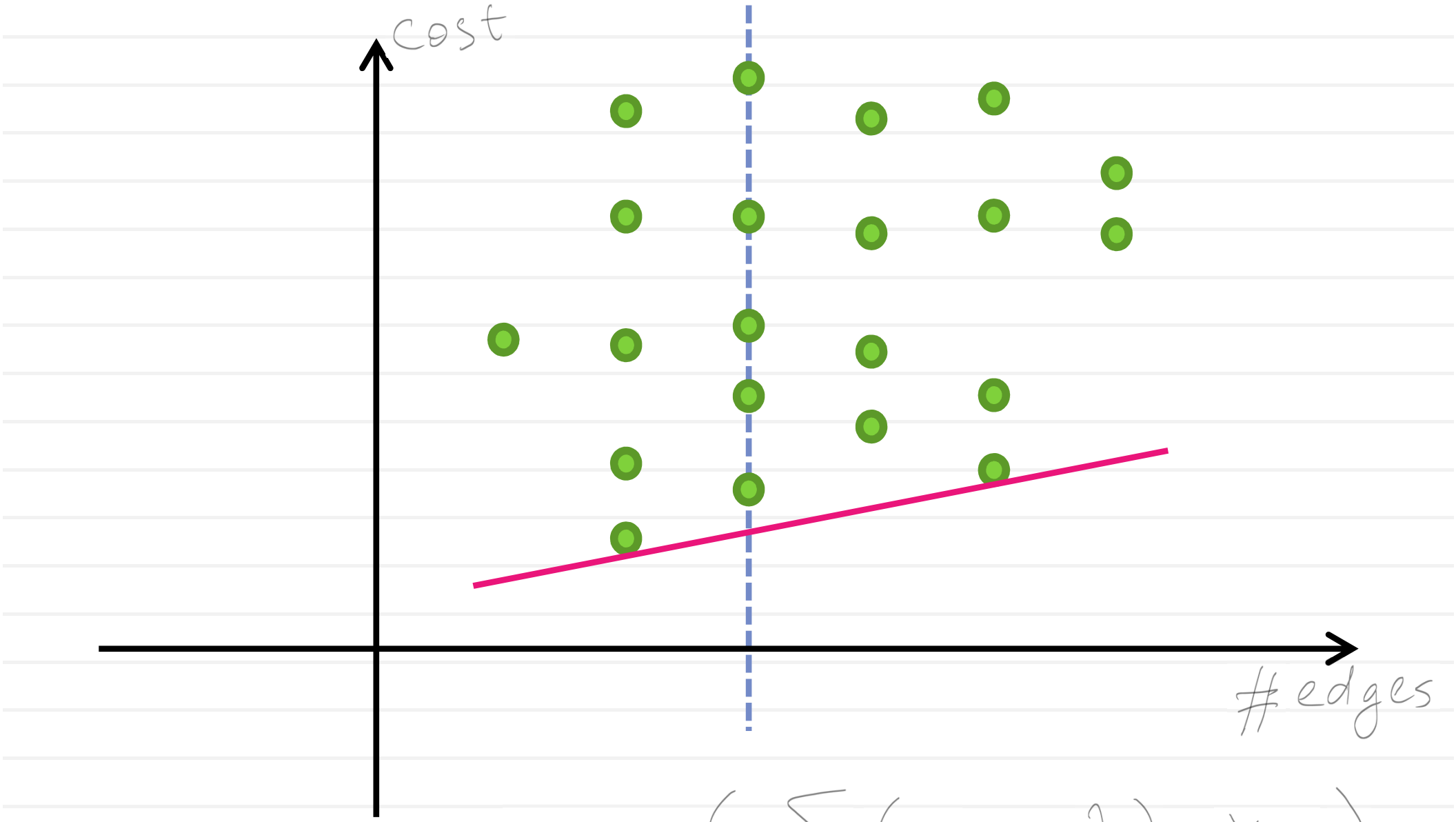
Case 1: strong duality holds



The shape of the dual function

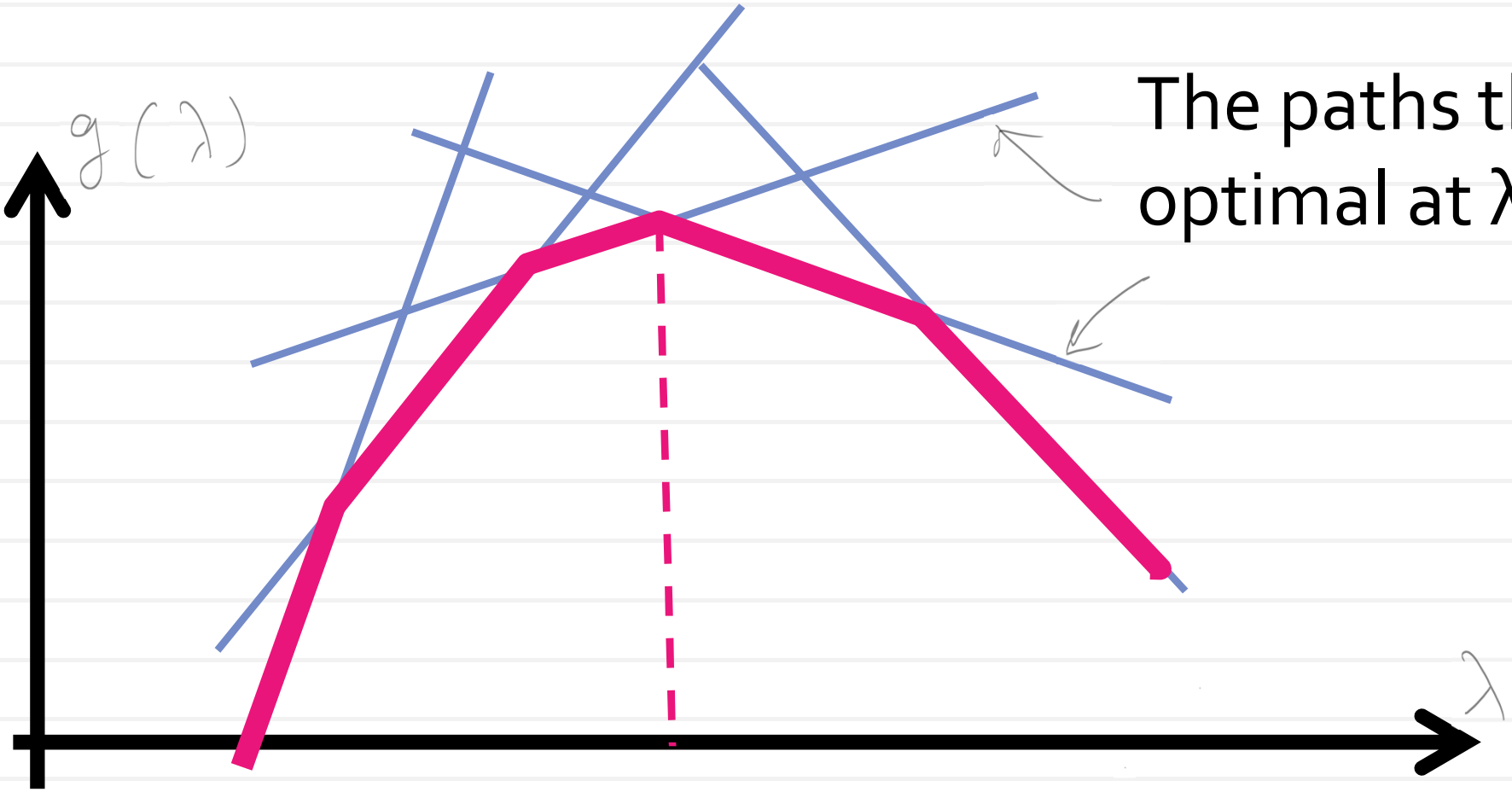


Case 2: strong duality does not holds



$$g(\lambda) = \lambda K + \min_{x \in Path(G)} \left(\sum (\omega_{ij} - \lambda) \cdot x_{ij} \right)$$

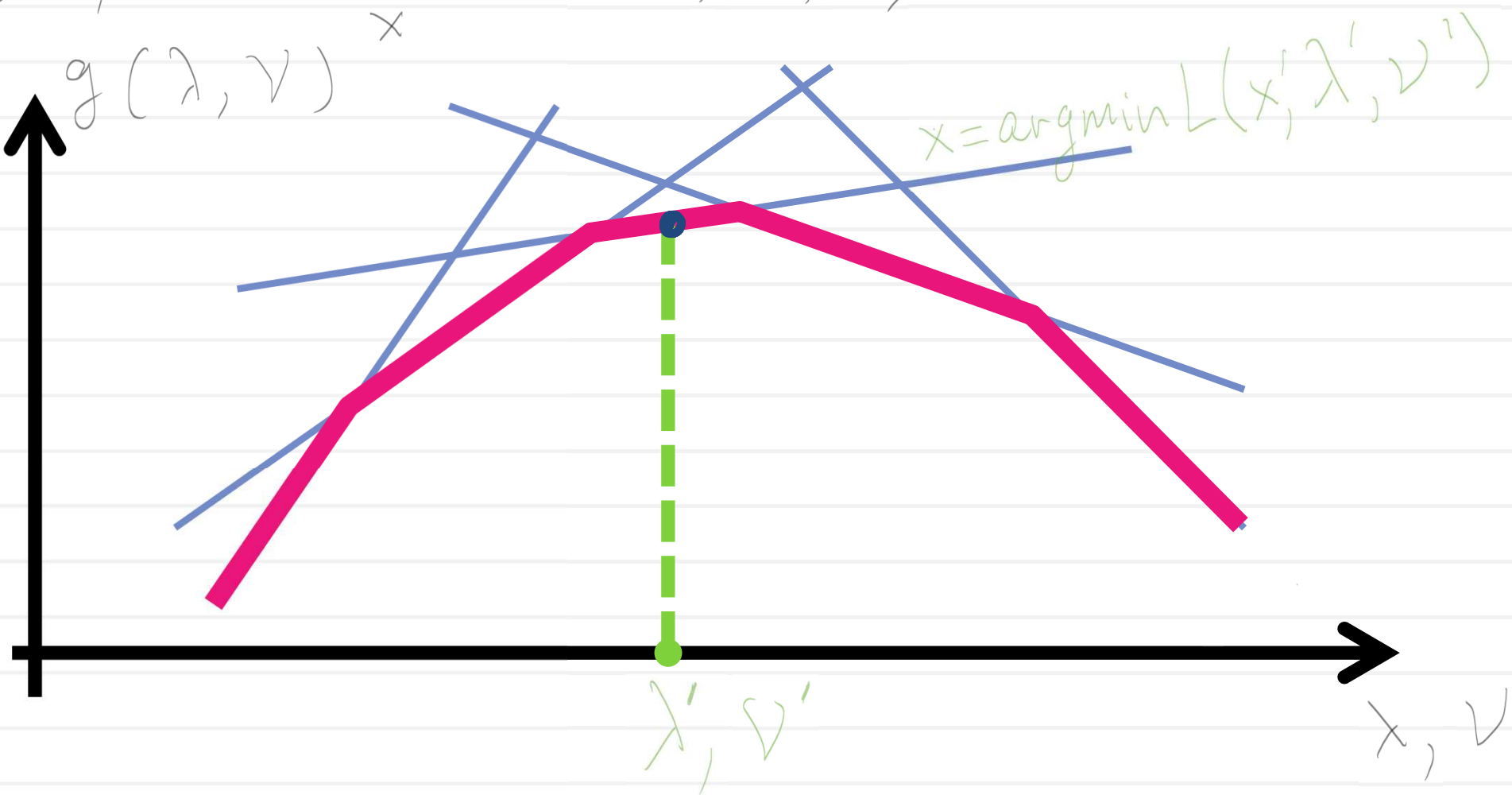
The shape of the dual function



The paths that are optimal at λ*

Recap

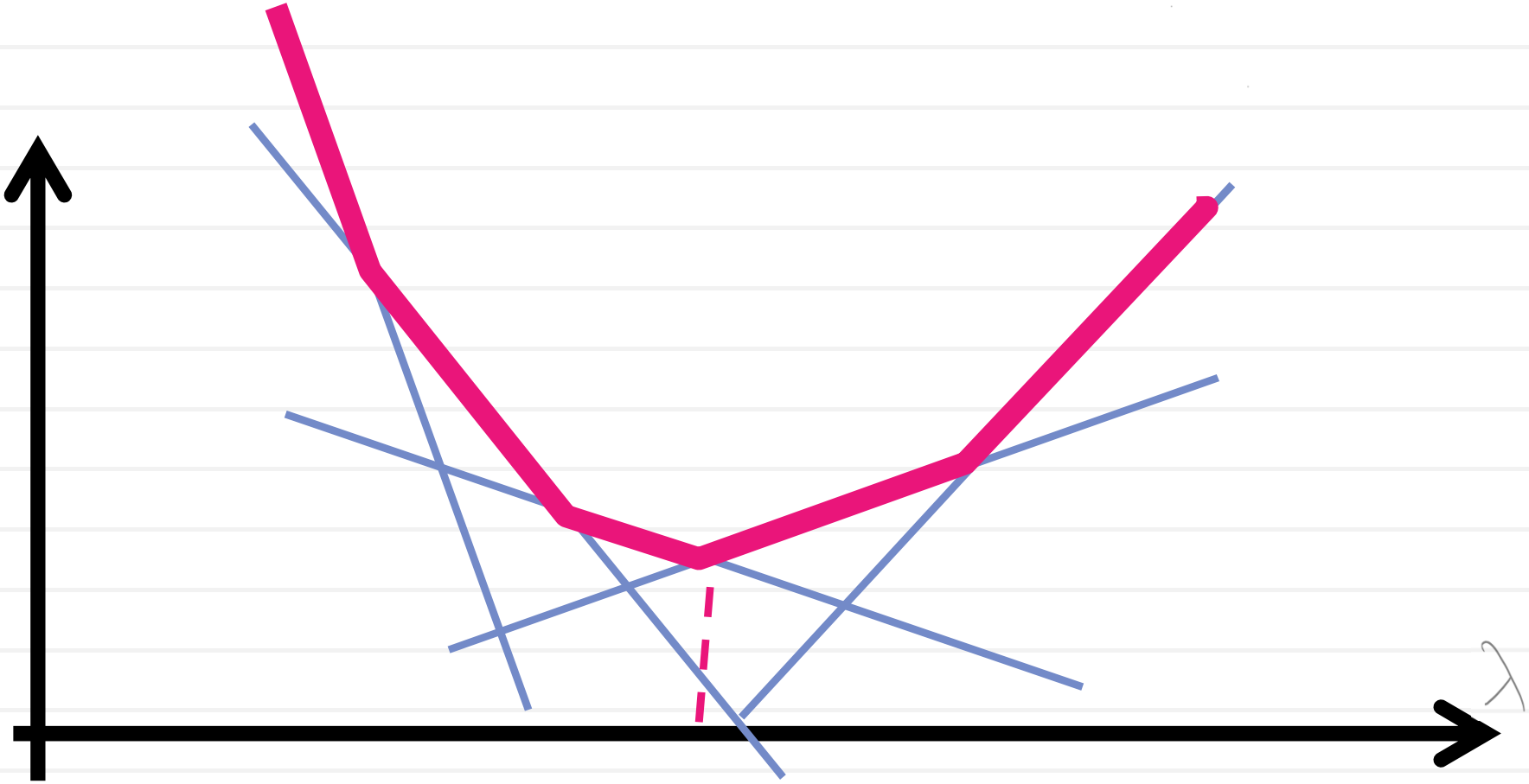
$$g(\lambda, \nu) = \min_x L(x, \lambda, \nu)$$



$$g(\lambda, \nu) \leq g(\lambda', \nu') + f(x')^T(\lambda - \lambda') + h(x')^T(\nu - \nu')$$

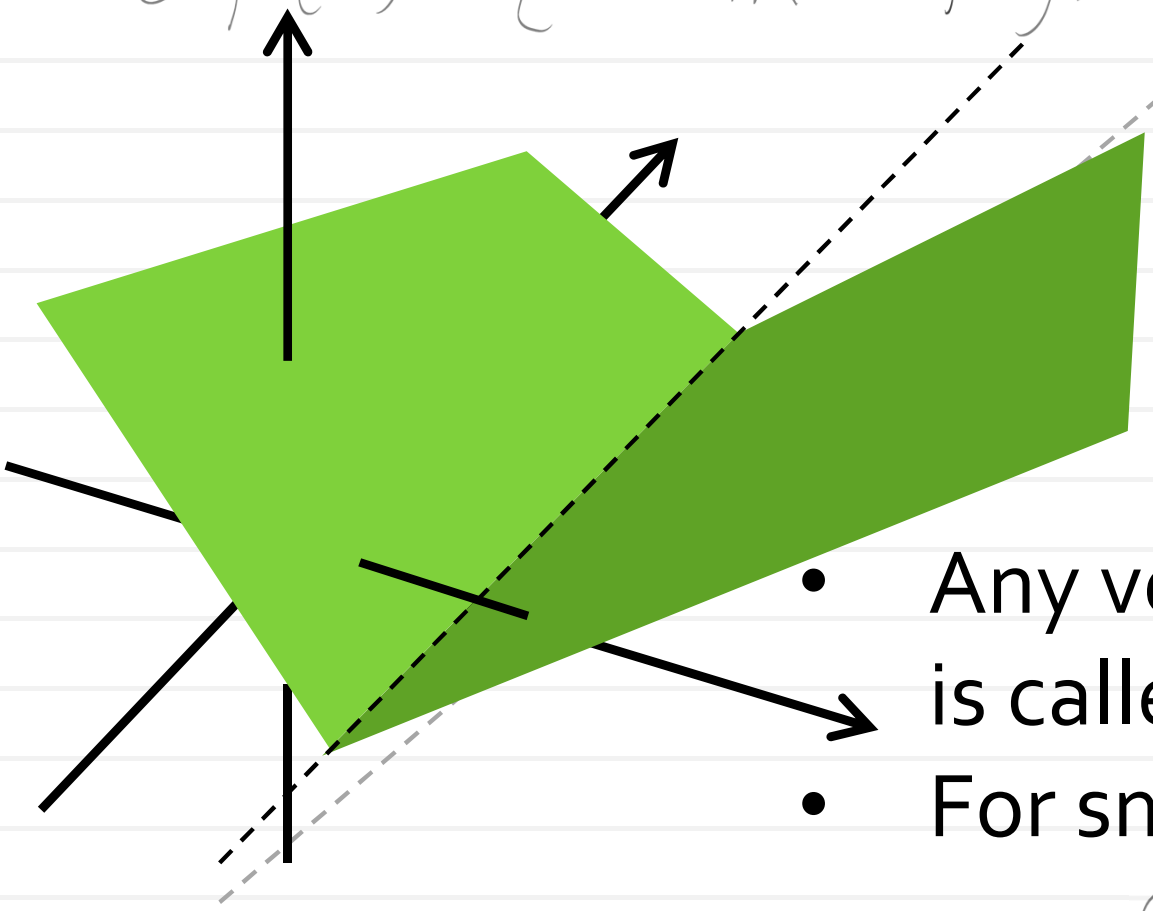
$$\nabla g = (f_i(x), h_i(x))$$ - *(sub)gradient of g*

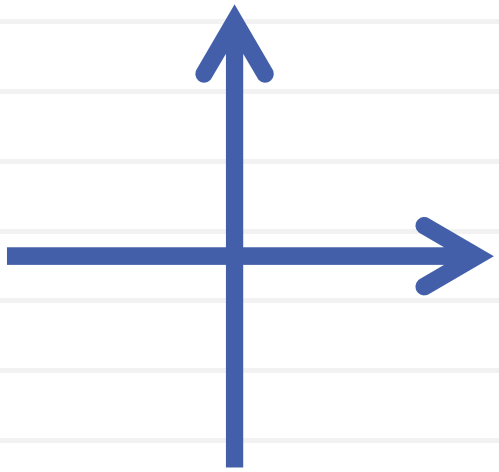
Minimizing non-differentiable convex functions



Subdifferential

$$\partial f(x) = \{ s \in \mathbb{R}^n : f(y) \geq f(x) + \langle s, y - x \rangle \}$$





- Any vector in subdifferential is called *subgradient*
- For smooth functions

$$\partial f(x) = \{ \nabla f(x) \}$$

- Convex functions have non-empty subdifferential (why?)

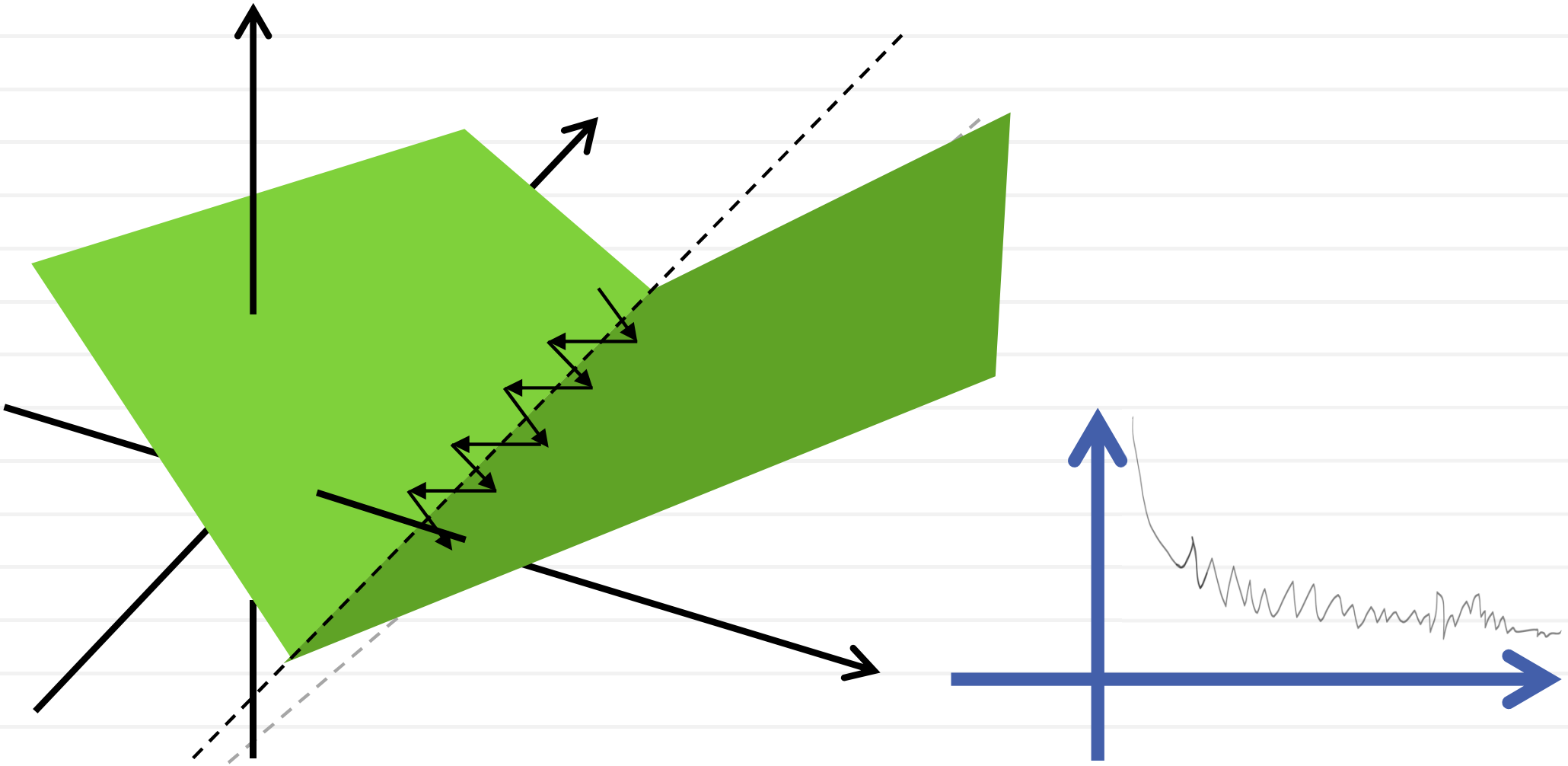
Subgradient descent

- A subgradient is not necessarily a descent direction
- However, persistence pays off:

```
for k = 1..K
        g = GetSubgradient(f,x(k))
        x(k+1) = x(k)-α_k g / ||g||
end
```

Corollary: if $\alpha_k \geq 0,\ \alpha_k \to 0,\ \sum \alpha_k \to \infty$ then $f(x(k)) \to \min_x f(x)$

Subgradient descent



- A subgradient is not necessarily a descent direction
- However, persistence pays off

Subgradient descent

```
xbest = x(0), fbest = f(x(0))
for k = 0..K
        g = GetSubgradient(f,x(k))
        x(k+1) = x(k)-1/(sqrt(k)+βk)*g
        if f(x(k+1)) < fbest
                fbest = f(x(k+1))
                xbest = x(k+1)
        end
end
return xbest
```

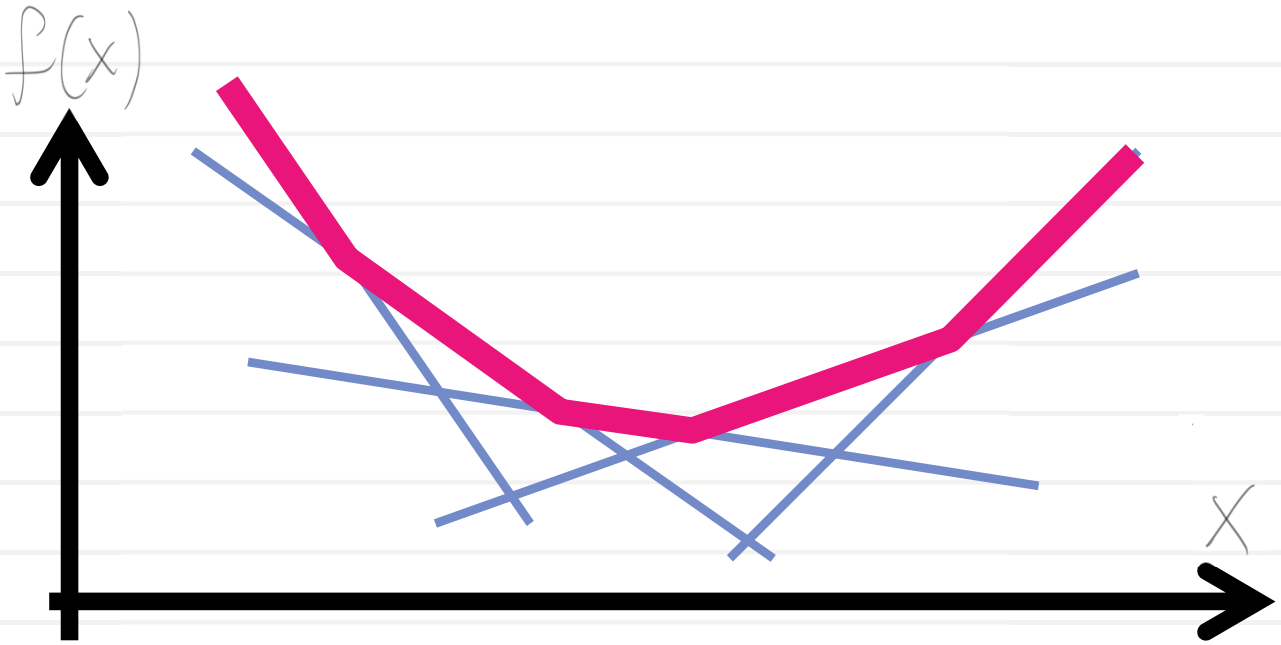
In practical implementation:

- "stashing" good solution is important
- Tuning schedule is important
- Even for a good schedule can be very slow

Cutting plane method

$$f(x) = \max_y \; f(y) + \tilde{\nabla} f(y)(x - y)$$

Minimizing f is equivalent to solving the LP:



$$\min_{x,\xi} \; \xi$$

$$\forall y: \quad \xi \geq f(y) + \tilde{\nabla} f(y)(x - y)$$

Often converges much faster!

Cutting plane method

$$f(x) = \max_y \; f(y) + \tilde{\nabla} f(y)(x - y)$$

$$\min_{x,\xi} \; \xi \qquad (*)$$
$$y \in Y \;\; \xi \geq f(y) + \tilde{\nabla} f(y)(x - y)$$

Initialize $Y = \{y1, y2, \ldots yn\}$

for k = 1..K

(y, ξ) = argmin (*)

Add y to Y

end

Cutting plane method

$$\min_{x,\xi} \; \xi \quad \text{s.t.} \quad y \in Y \;\; \xi \geq f(y) + \tilde{\nabla} f(y)(x - y) \qquad (*)$$

Initialize $Y = \{y1, y2, \ldots yn\}$

for k = 1..K

(y, ξ) = argmin (*)

Add y to Y

end

Problems:

- LPs grows and become too slow and too big
- The optimal points oscillate wildly ("zigzaging")

Bundle method

$$\min_{x,\xi} \ \xi + \lambda (x - x(k))^2 \qquad (*)$$

$$y \in Y \ \} \geq f(y) + \tilde{\nabla} f(y)(x - y)$$

- Penalize deviation from the last solution
- The sequence of solutions changes slowly
- As with trust-region methods, can tweak λ on the fly
- Keep only the last T points in Y (limits time/memory)
- The devil is in details (but "black box" implementations exist)

Plan for the second part

- What happens when we consider the primal and dual problems together?
- What do dual variables tell us about the primal program?
- What does duality mean for "smooth" problems?

Duality gap and certificates of optimality

Let x to be any feasible primal set of variables, and (λ, ν) be any feasible dual set of variables (i.e. $\lambda \geq o$). Then:

$$g(\lambda,\nu) \leq g(\lambda^*,\nu^*) \leq f_o(x^*) \leq f_o(x)$$

duality gap (x,λ,ν)

$$f_o(x) - f_o(x^*) \leq f_o(x) - g(\lambda,\nu) = gap(x,\lambda,\nu)$$

Any feasible (x, λ, ν) give us a *certificate* on how far is $f_o(x)$ from optimality.

Strong duality and saddle point

Assume the strong duality holds (and optimal values are attained):

$$g(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) = f_0(x^*)$$

$$\forall x \quad L(x, \lambda^*, \nu^*) \geq g(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*)$$

$$x^* = \operatorname{argmin}_x L(x, \lambda^*, \nu^*)$$

$$\forall \lambda \geq 0, \nu \quad L(x^*, \lambda, \nu) \leq f_0(x^*) = L(x^*, \lambda^*, \nu^*)$$

$$(\lambda^*, \nu^*) = \operatorname{arg\,max}_{\lambda \geq 0, \nu} L(x^*, \lambda, \nu)$$

Thus (x^*, λ^*, ν^*) correspond to a saddle point of the Lagrangian.

Duality for smooth functions

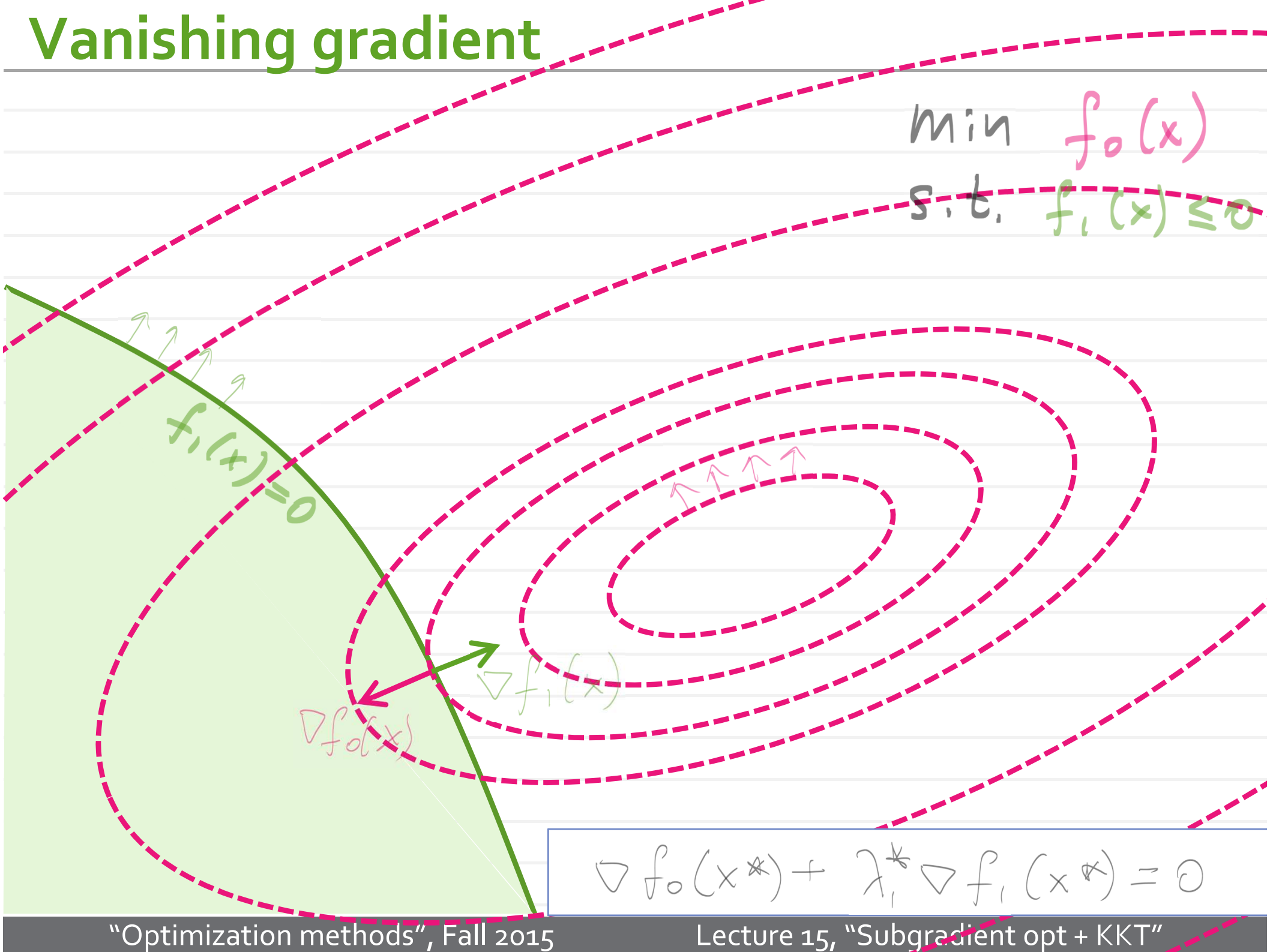
Let us assume that f_i, h_i are differentiable and strong duality holds.

We know that $x^* = \operatorname{argmin}_x L(x, \lambda^*, \nu^*)$

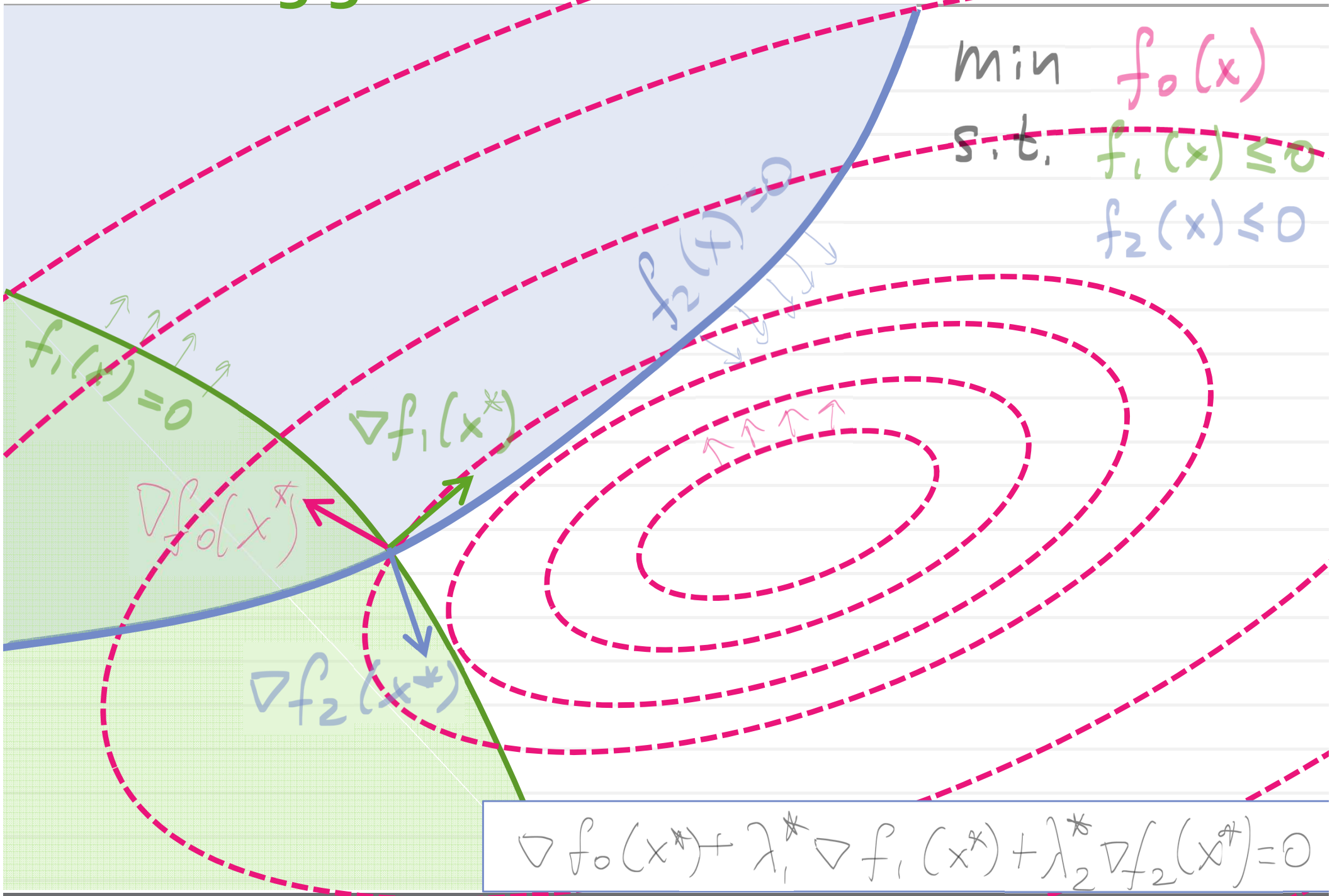
Then the gradient of the Lagrangian w.r.t. x must vanish: $\frac{\partial L}{\partial x}(x^*, \lambda^*, \nu^*) = 0$

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0$$

Vanishing gradient



Vanishing gradient



Complementary slackness

Assume the strong duality holds (and optimal values are attained):

$$L(x^*, \lambda^*, \nu^*) = f_0(x^*)$$

$$f_0(x^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*)$$

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0 \qquad \sum_{i=1}^{p} \nu_i^* h_i(x^*) = 0$$

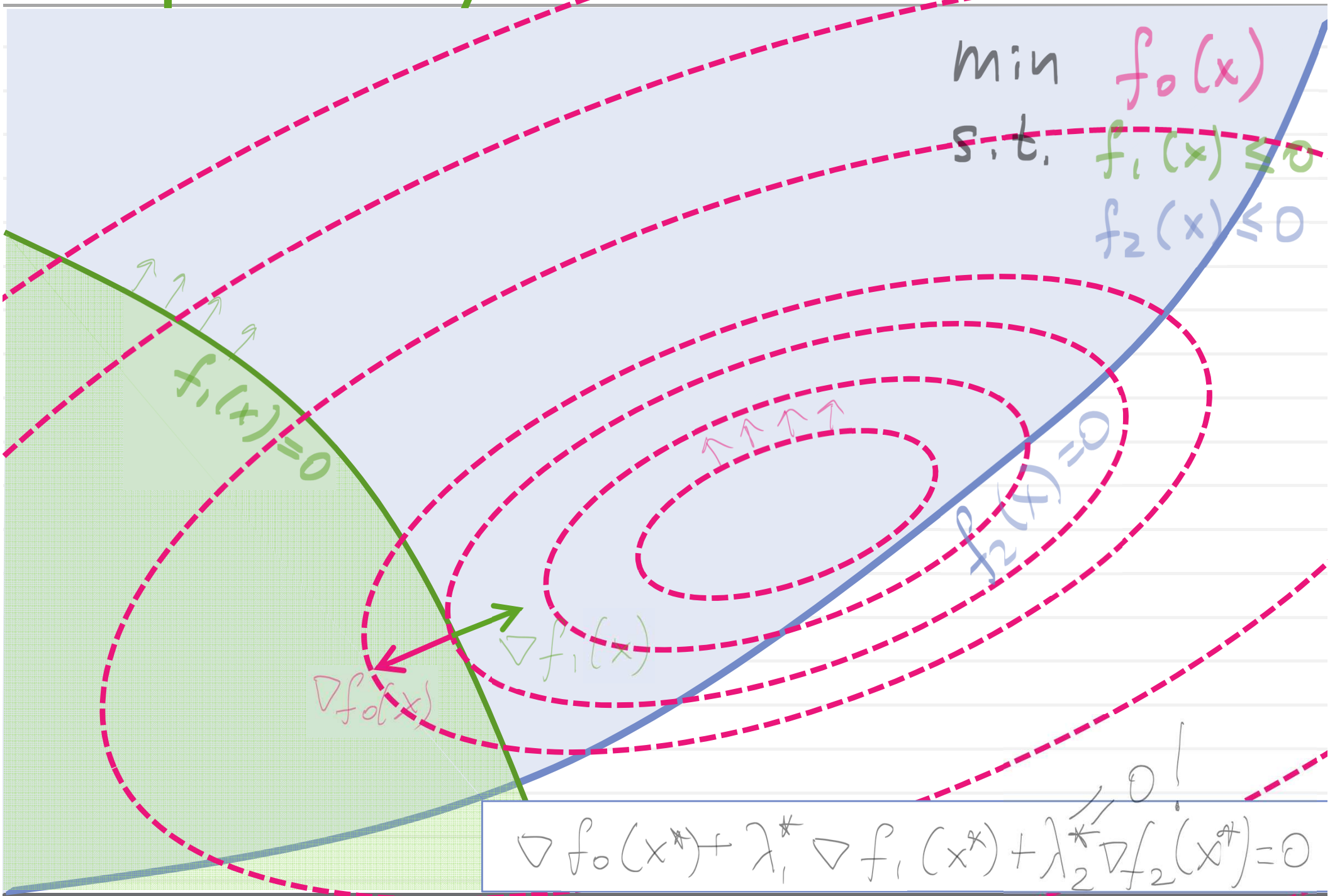
Since $\lambda_i^* \geq 0 \quad f_i(x^*) \leq 0$

We have: $\forall i \quad \lambda_i^* f_i(x^*) \leq 0$

Since they all sum to 0, they all should equal 0.

Complementary slackness $\forall i \quad \lambda_i^* f_i(x^*) = 0$

Complementary slackness



Karush-Kuhn-Tucker conditions

$$\min_x \; f_0(x)$$
$$\text{s.t.:} \quad f_i(x) \le 0$$
$$h_i(x) = 0$$

If (x^*, λ^*, ν^*) are a set of optimal primal and dual variables and strong duality holds, then the following must be true:

$$\begin{cases} f_i(x^*) \le 0 \quad i = 1 \dots m \qquad h_i(x^*) = 0 \quad i = 1 \dots p \\ \lambda_i^* \ge 0 \quad i = 1 \dots m \\ \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0 \\ \lambda_i^* f_i(x^*) = 0 \quad i = 1 \dots m \end{cases}$$

Karush-Kuhn-Tucker conditions

$$
\begin{cases}
f_i(x^*) \le 0 \quad i = 1 \dots m \qquad h_i(x^*) = 0 \quad i = 1 \dots p \\
\lambda_i^* \ge 0 \quad i = 1 \dots m \\
\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0 \\
\lambda_i^* f_i(x^*) = 0 \quad i = 1 \dots m
\end{cases}
$$

Sufficiency of KKT

Theorem: if the primal problem is convex and (x', λ', ν') are a set of optimal primal and dual variables satisying KKT, then they are primal and dual optimal.

Proof:

$$x' = \operatorname{argmin}_x L(x, \lambda', \nu')$$

$$g(\lambda', \nu') = L(x', \lambda', \nu') =$$

$$= f_0(x') + \sum \lambda_i' f_i(x') + \sum \nu_i' h_i(x') =$$

$$= f_0(x')$$

Example 1: unconstrained optimization

$$\min_x \; f_0(x)$$

KKT are turned into:

$$\nabla f_0(x) = 0$$

We get a Newton method.

KKT = "primal feasibility" + "dual feasibility"+ "vanishing gradient" + "complementary slackness"

Equality-constrained quadratic program

$$\min \frac{1}{2} x^T Q x + p^T x$$

$$s.t.: \quad Ax = b$$

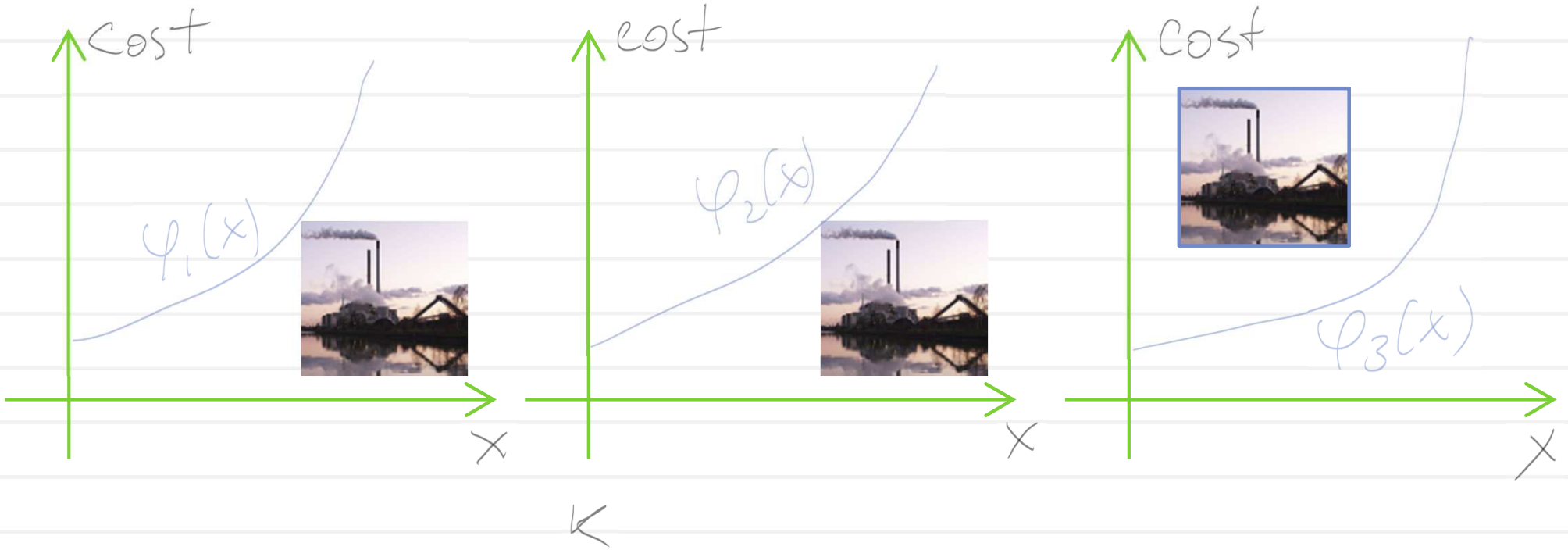
KKT results in:

M+N equations on M+N variables

$$\begin{cases} Qx + p + A^T \nu = 0 \\ Ax = b \end{cases}$$

KKT = "primal feasibility" + "dual feasibility"+ "vanishing gradient" + "complementary slackness"

Economic dispatch



$$\min_{x} \sum_{i=1} \varphi_i(x_i)$$

Total cost

$$s.t.: \sum x_i = D$$

Meet the demand

Economic dispatch

$$\min_{x} \sum_{i=1}^{K} \varphi_i(x_i)$$

$$s.t.: \sum x_i = D$$

Note: we denote the dual variable as λ because of the traditional notation for this particular application (should have been ν).

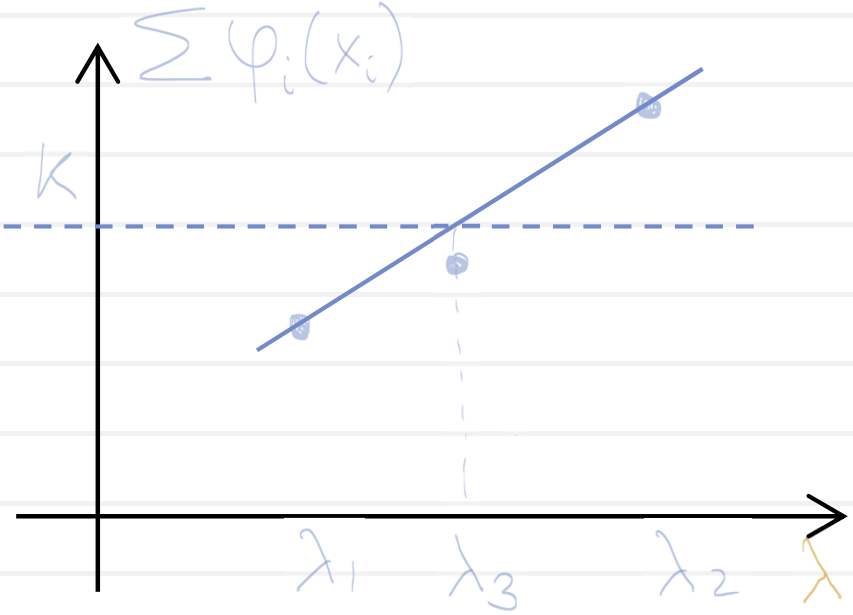
KKT results in:

$$\begin{cases} \sum \nabla \varphi_i(x) - \begin{pmatrix} \lambda \\ \lambda \\ \vdots \\ \lambda \end{pmatrix} = 0 \\ \sum x_i = D \end{cases} \Rightarrow \begin{cases} \varphi_i''(x_i) = \lambda \\ \sum x_i = D \end{cases}$$

KKT = "primal feasibility" + "dual feasibility"+ "vanishing gradient" + "complementary slackness"

Economic dispatch: lambda search

$$\begin{cases} \varphi_i'(x_i) = \lambda \\ \sum x_i = D \end{cases}$$



Inequality constrained QP

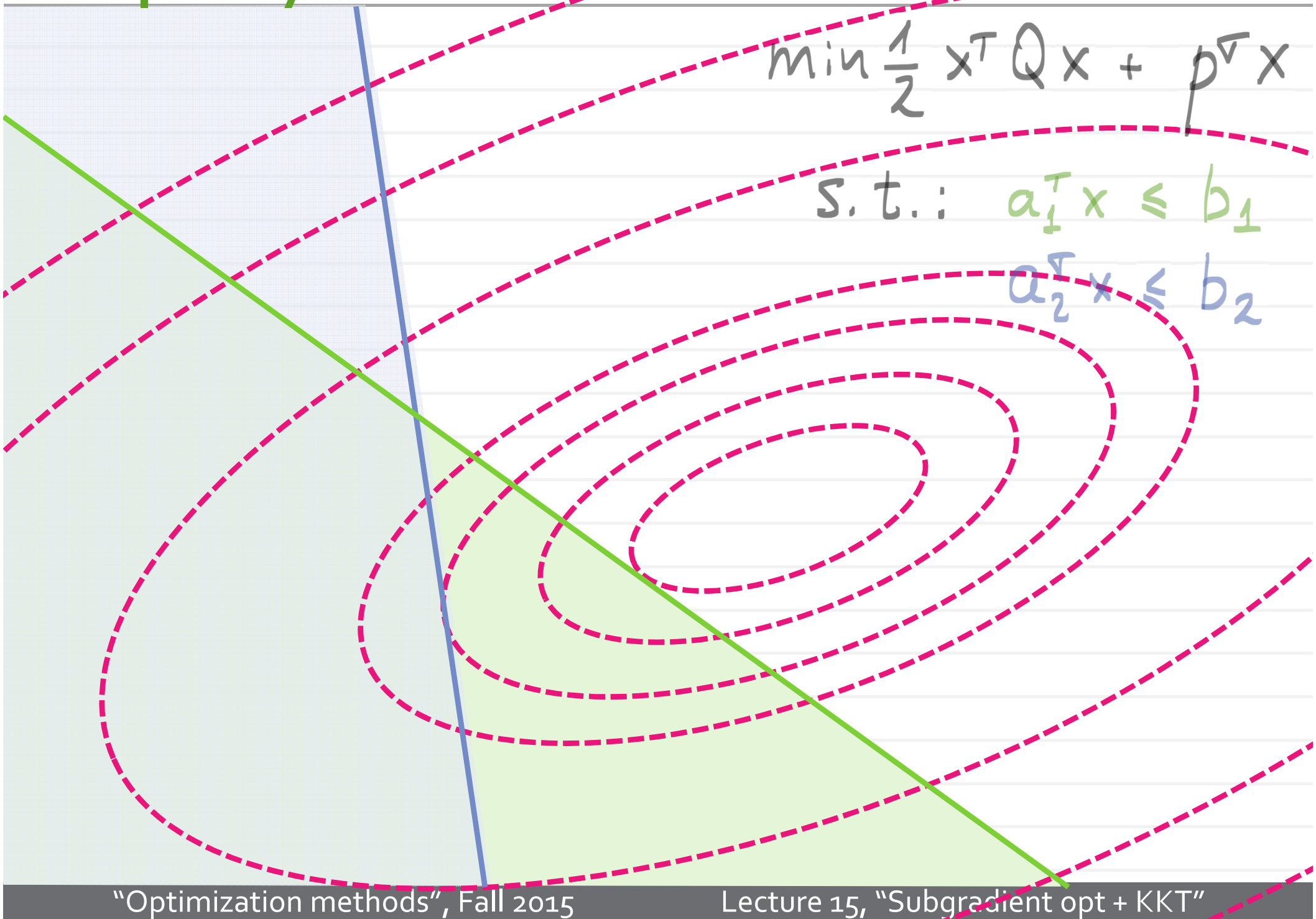
$$\min \frac{1}{2} x^T Q x + p^T x$$

$$\text{s.t.:} \quad a_1^T x \leq b_1$$

$$a_2^T x \leq b_2$$

$$\begin{cases} Qx + p + a_1 \lambda_1 + a_2 \lambda_2 = 0 \\ a_1^T x \leq b_1 \qquad \lambda_1 \geq 0 \\ a_2^T x \leq b_2 \qquad \lambda_2 \geq 0 \\ \lambda_1 (a_1^T x - b_1) = 0 \qquad \lambda_2 (a_2^T x - b_2) = 0 \end{cases}$$

Inequality constrained QP



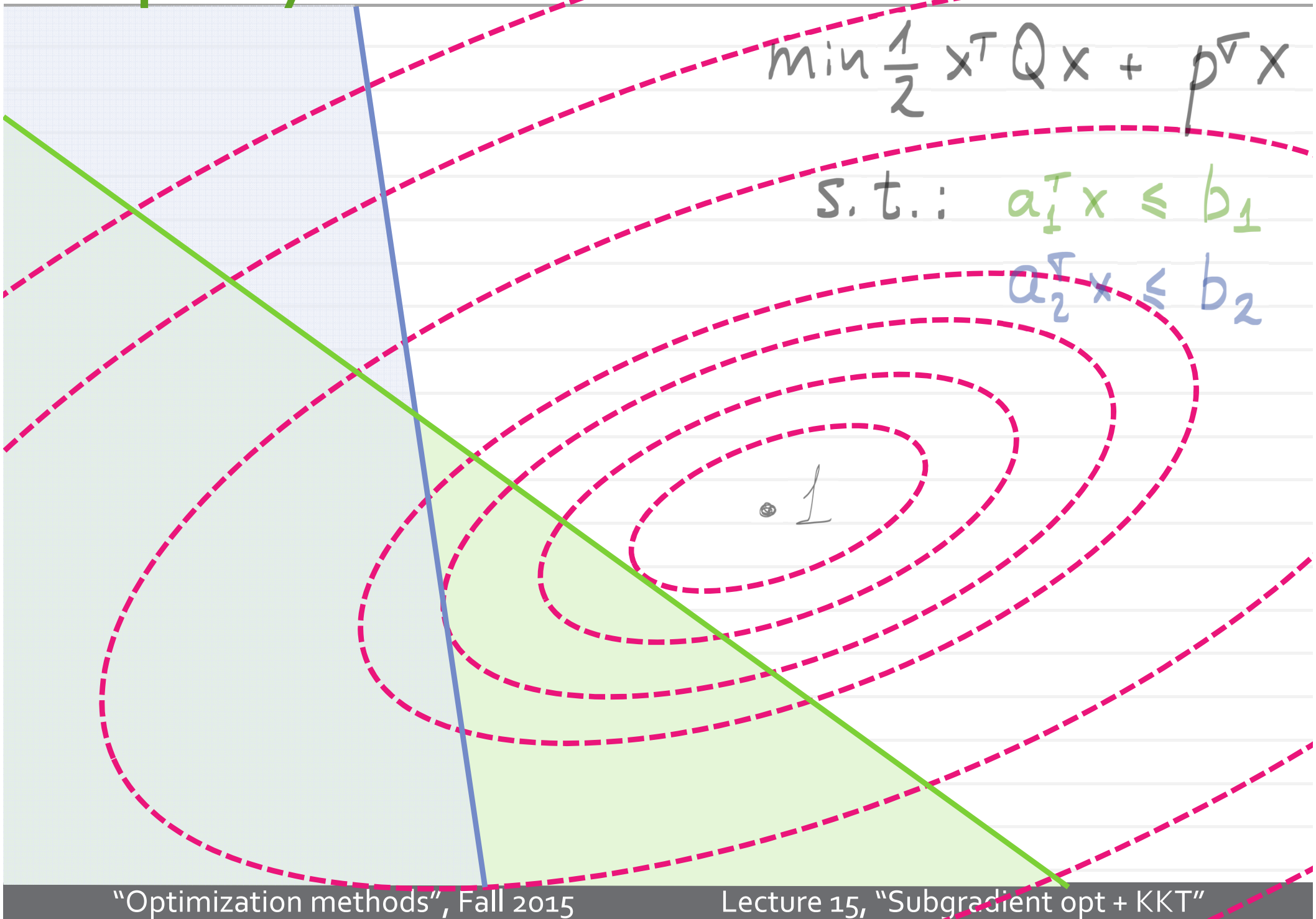
Inequality constrained QP

$$\begin{cases} Qx + p + a_1 \lambda_1 + a_2 \lambda_2 = 0 \\ a_1^T x \le b_1 \qquad \lambda_1 \ge 0 \\ a_2^T x \le b_2 \qquad \lambda_2 \ge 0 \\ \lambda_1 (a_1^T x - b_1) = 0 \qquad \lambda_2 (a_2^T x - b_2) = 0 \end{cases}$$

Case 1: $\lambda_1 = 0 \quad \lambda_2 = 0$

$$\begin{cases} Qx + p = 0 \\ a_1^T x \le b_1 \\ a_2^T x \le b_2 \end{cases}$$

Inequality constrained QP



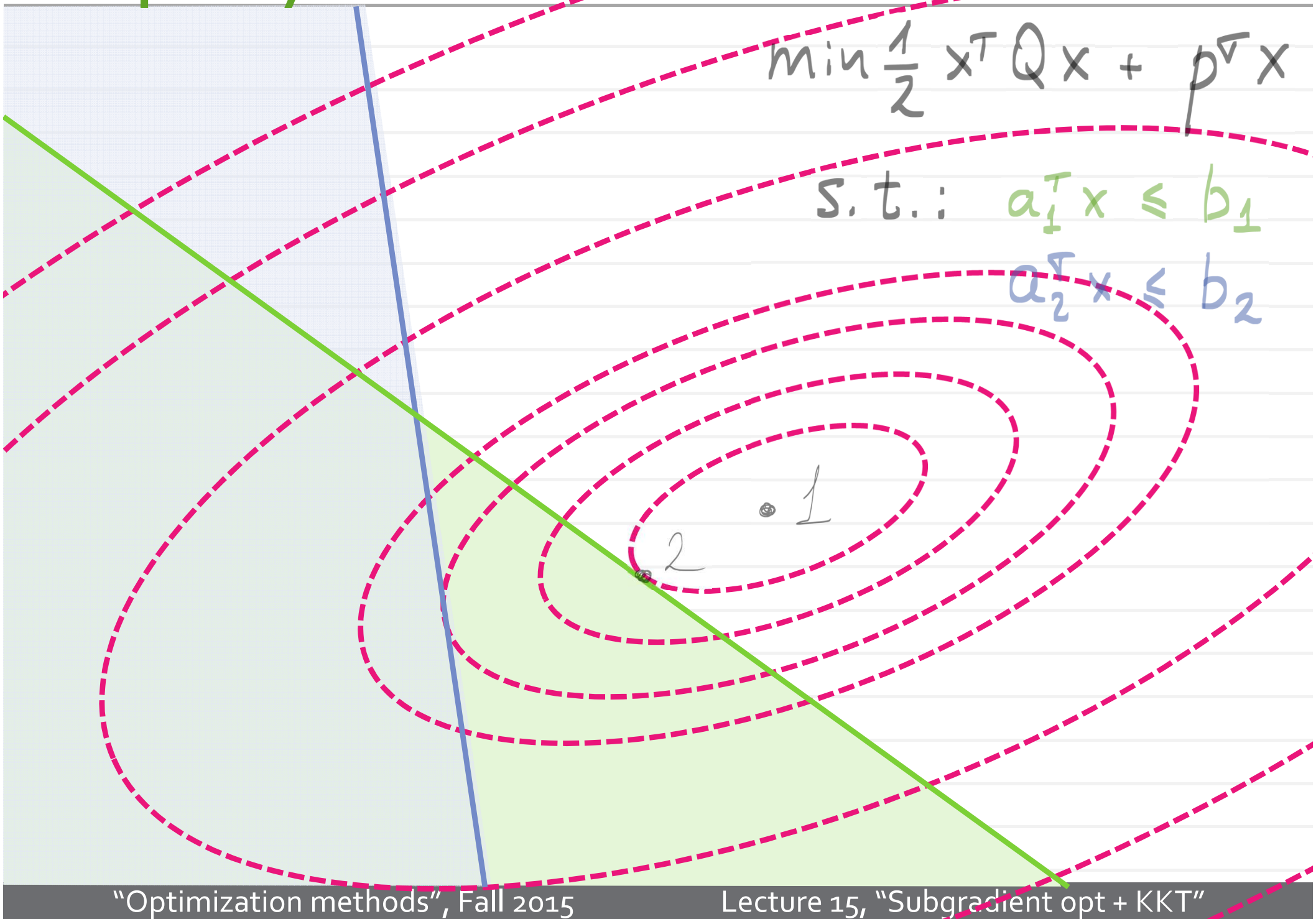
Inequality constrained QP

$$\begin{cases} Qx + p + a_1\lambda_1 + a_2\lambda_2 = 0 \\ a_1^T x \le b_1 \qquad \lambda_1 \ge 0 \\ a_2^T x \le b_2 \qquad \lambda_2 \ge 0 \\ \lambda_1(a_1^T x - b_1) = 0 \qquad \lambda_2(a_2^T x - b_2) = 0 \end{cases}$$

Case 2: $\lambda_1 > 0$, $\lambda_2 = 0$

$$\begin{cases} Qx + p + a_1\lambda_1 = 0 \\ a_1^T x = b_1 \\ a_2^T x \le b_2 \qquad \lambda_1 > 0 \end{cases}$$

Inequality constrained QP



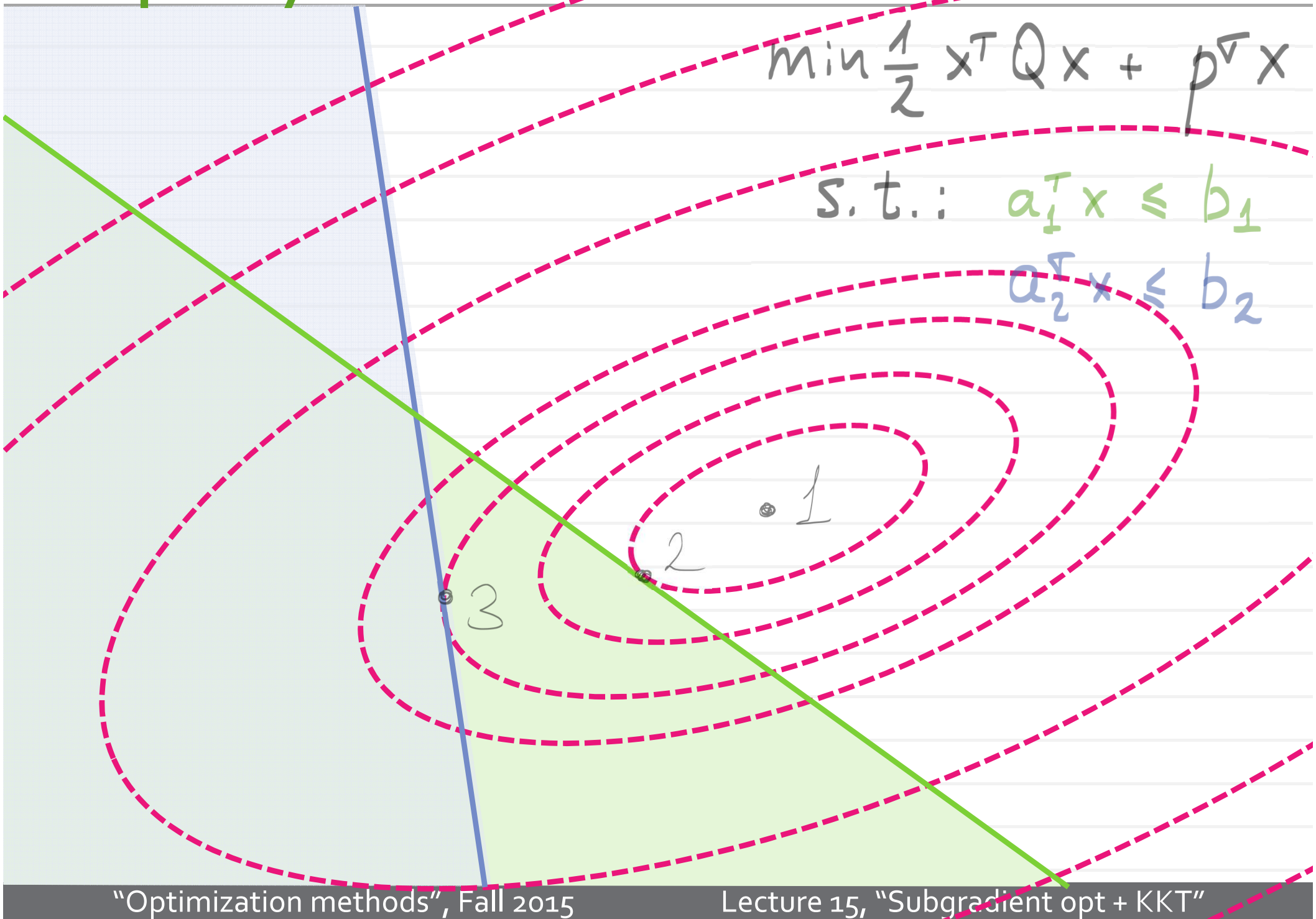
Inequality constrained QP

$$\begin{cases} Qx + p + a_1\lambda_1 + a_2\lambda_2 = 0 \\ a_1^T x \le b_1 \qquad \lambda_1 \ge 0 \\ a_2^T x \le b_2 \qquad \lambda_2 \ge 0 \\ \lambda_1(a_1^T x - b_1) = 0 \qquad \lambda_2(a_2^T x - b_2) = 0 \end{cases}$$

Case 3: $\lambda_1 = 0$, $\lambda_2 > 0$

$$\begin{cases} Qx + p + a_2\lambda_2 = 0 \\ a_2^T x = b_2 \\ a_1^T x \le b_1 \qquad \lambda_2 > 0 \end{cases}$$

Inequality constrained QP



Inequality constrained QP

$$\begin{cases} Qx + p + a_1\lambda_1 + a_2\lambda_2 = 0 \\ a_1^T x \le b_1 \qquad \lambda_1 \ge 0 \\ a_2^T x \le b_2 \qquad \lambda_2 \ge 0 \\ \lambda_1(a_1^T x - b_1) = 0 \qquad \lambda_2(a_2^T x - b_2) = 0 \end{cases}$$

Case 4: $\lambda_1 > 0 \quad \lambda_2 > 0$

$$\begin{cases} Qx + p + a_1\lambda_1 + a_2\lambda_2 = 0 \\ a_1^T x = b_1 \qquad \lambda_1 > 0 \quad \lambda_2 > 0 \\ a_2^T x = b_2 \end{cases}$$

Inequality constrained QP