Lec 13 - Math Basis for Rate Distortion Optimization

Zhu Li

Dept of CSEE, UMKC

Office: FH560E, Email: lizhu@umkc.edu, Ph: x 2346.

http://l.web.umkc.edu/lizhu

slides created with WPS Office Linux and EqualX LaTex equation editor

Outline

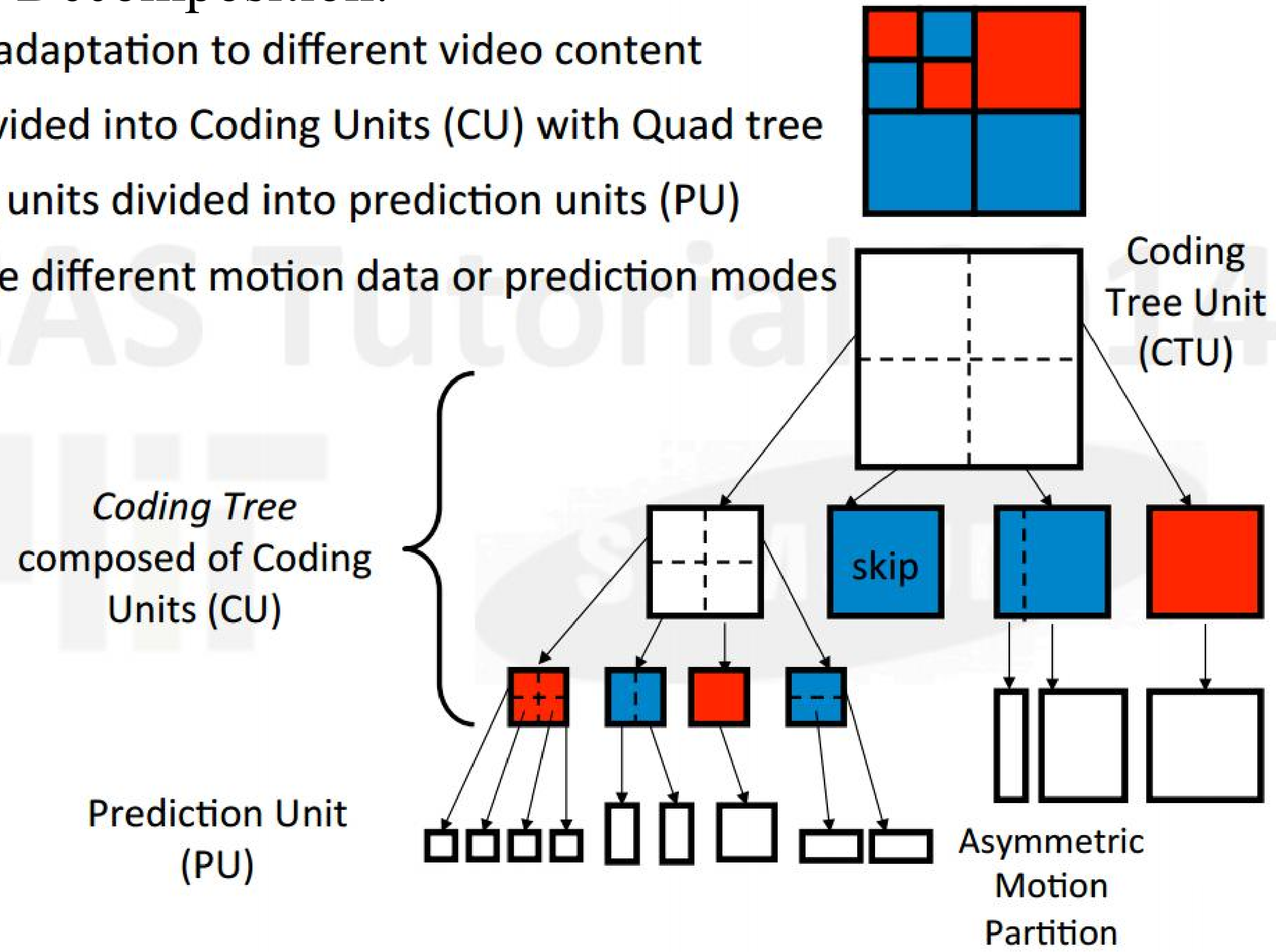
- ❑ Recap of HEVC system
- ❑ R-D Optimization Basis
- ❑ Summary

HEVC Coding Structure

Slide Credit: Vivienne Sze & Madhukar Budagavi, ISCAS 2014

❑ Quad Tree Decomposition:

- Better adaptation to different video content
- CTU divided into Coding Units (CU) with Quad tree
- Coding units divided into prediction units (PU)
- PU have different motion data or prediction modes



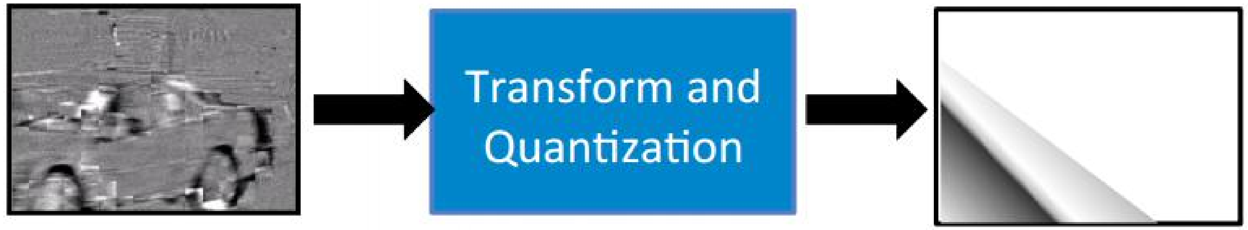
Ref:

G. Schuster, PhD Thesis, 1996: Optimal Allocation of Bits Among Motion, Segmentation and Residual

HEVC Transforms

❑ Transform + Quant:

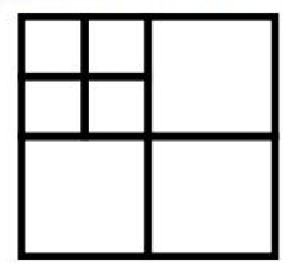
many pixels



few coefficients

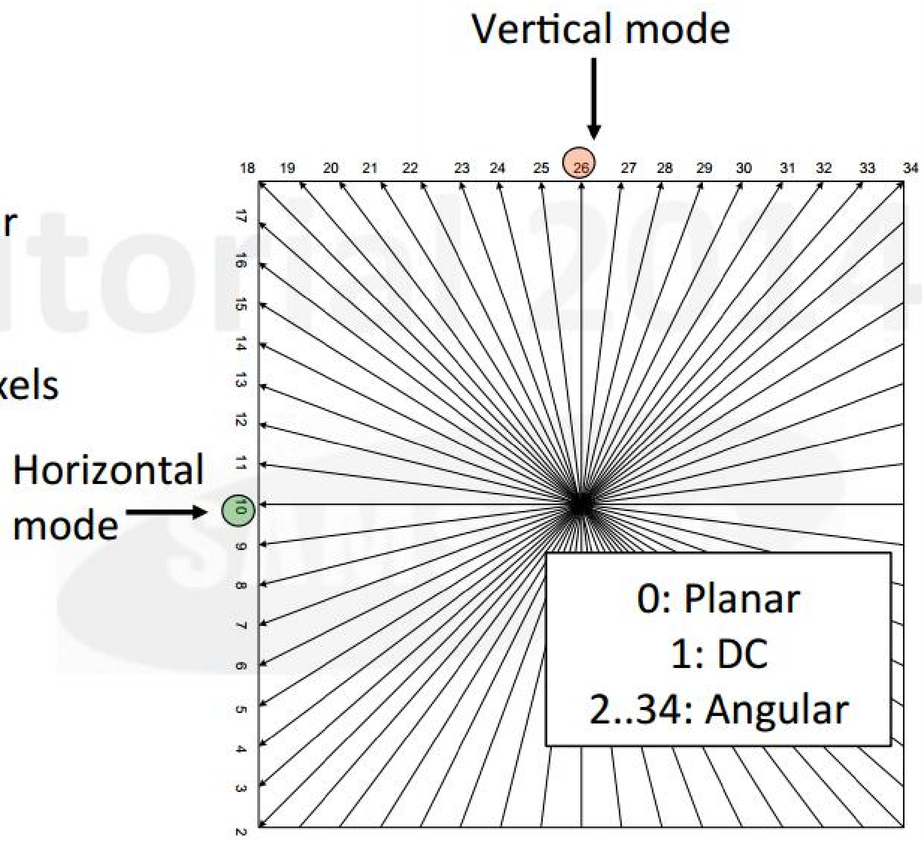
- HEVC supports 4x4, 8x8, 16x16, 32x32 integer transforms
  - Two types of 4x4 transforms (IDST-based for Intra, IDCT-based for Inter); IDCT-based transform for 8x8, 16x16, 32x32 block sizes
  - Integer transform avoids encoder-decoder mismatch and drift caused by slightly different floating point representations.
  - Parallel friendly matrix multiplication/partial butterfly implementation
  - Transform size signaled using Residual Quad Tree
- Achieves 5 to 10% increase in coding efficiency
- Increased complexity compared to H.264/AVC
  - 8x more computations per coefficient
  - 16x larger transpose memory

Represent residual of CU with TU quad tree



❑ Intra-Prediction Modes

- H.264/AVC has 10 modes
  - angular (8 modes), DC, planar
- HEVC has 35 modes
  - angular (33 modes), DC, planar
- Angular prediction
  - Interpolate from reference pixels at locations based on angle
- DC
  - Constant value which is an average of neighboring pixels (reference samples)
- Planar
  - Average of horizontal and vertical prediction



Intra-Predicted Basis

- ❑ As if it is a 1-non zero coefficient transform…

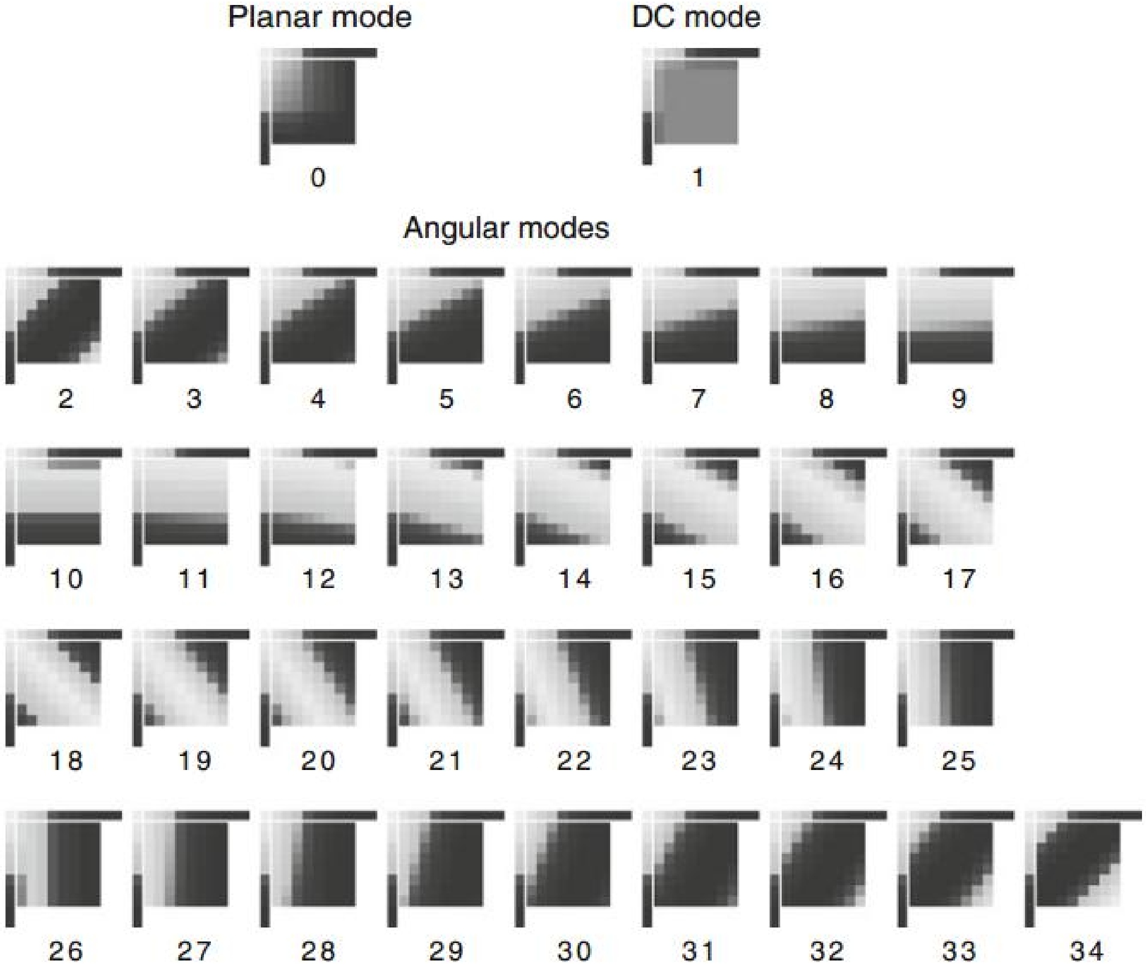
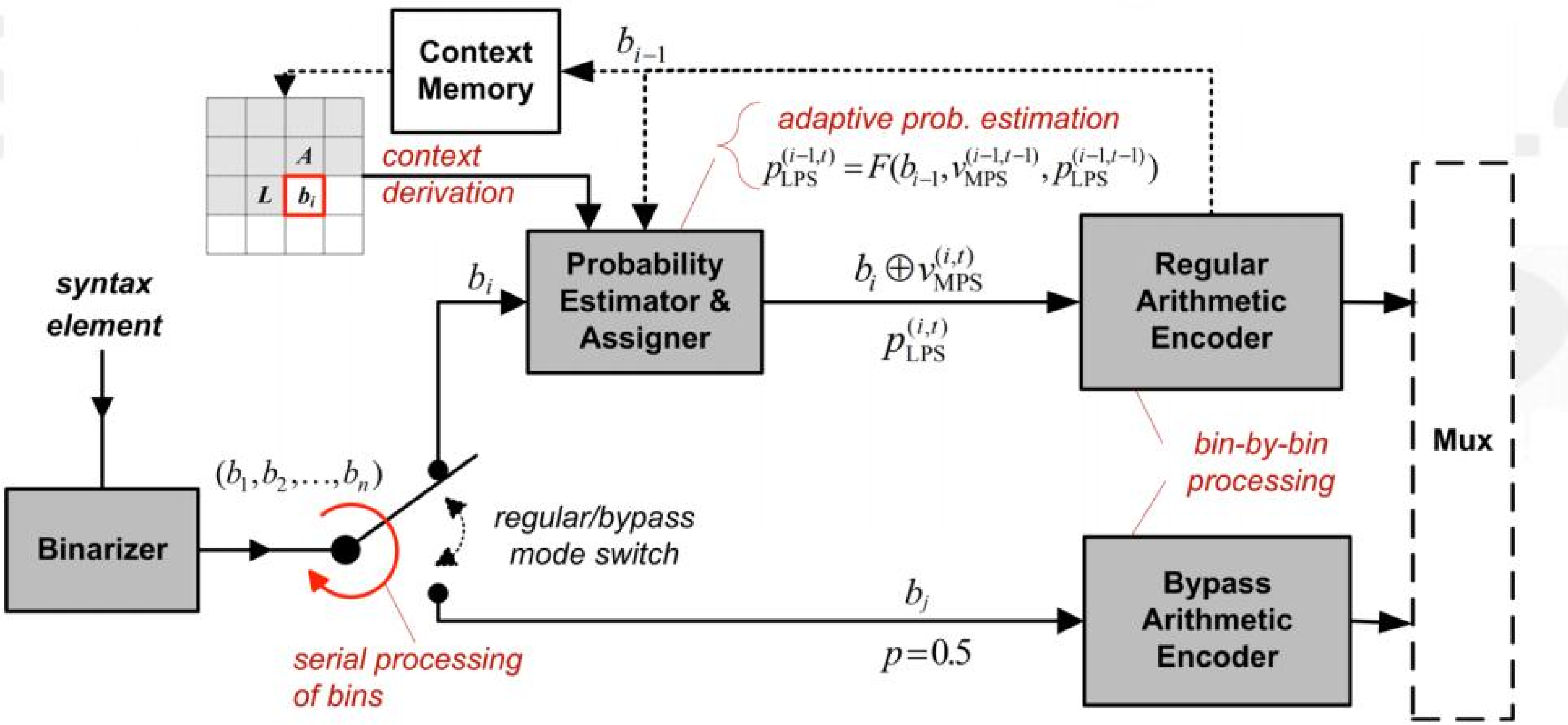


Fig. 4.2 Examples of 8 × 8 luma prediction blocks generated with all the HEVC intra prediction modes. Effects of the prediction post-processing can be seen on the top and left borders of the DC prediction (mode 1), top border of horizontal mode 10 and left border of vertical mode 26

Ref:
J. Laniema and W.-J. Han, "Intra Picture Prediction in HEVC", Chapter in, Springer-Velag Book on High Efficiency Video Coding (HEVC): Algorithms and Architectures, Springer, 2014. Ed. V. Sze et. Al.

❑ Binary Arithmetic Coding:

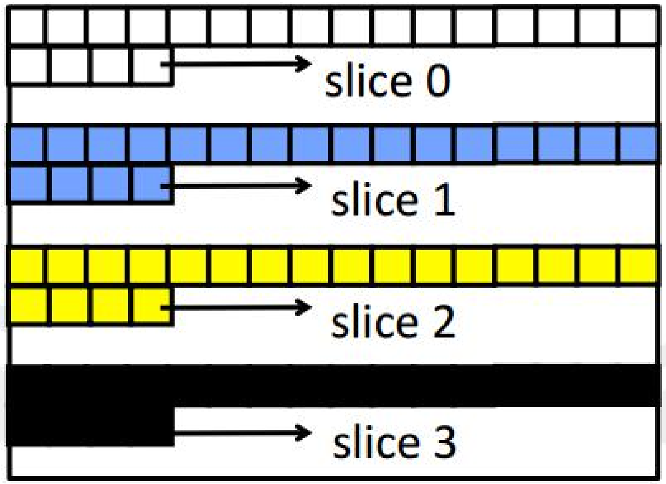
- HEVC uses Context Adaptive Binary Arithmetic Coding (CABAC)
  - 10 to 15% higher coding efficiency compared to CAVLC

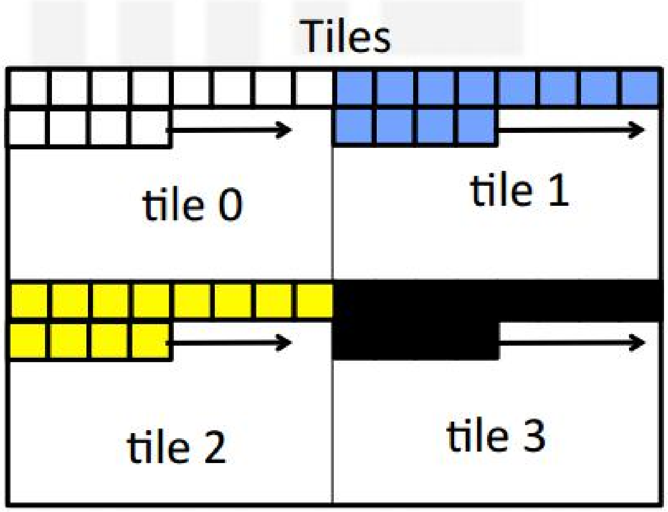


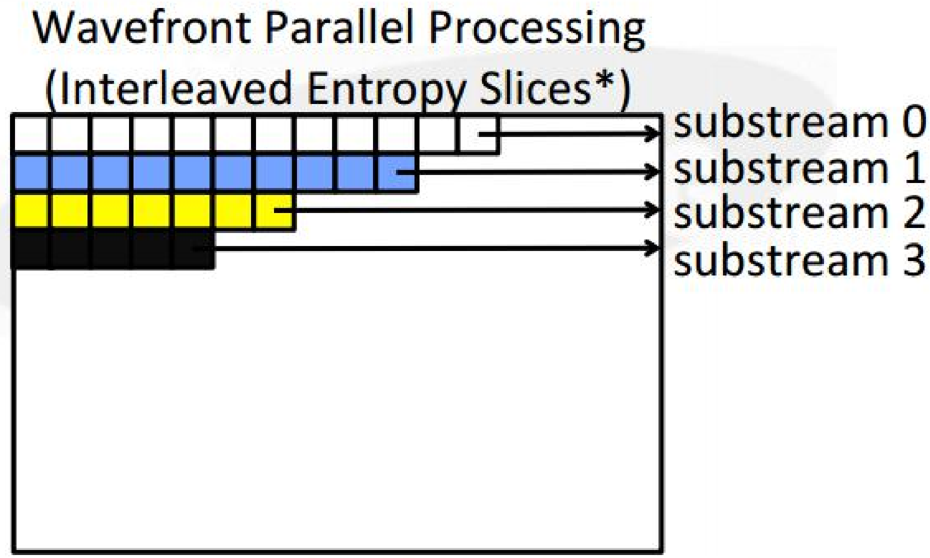
V. Sze, D. Marpe, "Entropy Coding in HEVC," *High Efficiency Video Coding (HEVC): Algorithms and Architectures*, Springer, 2014.

Parallel Processing Tools: Slice/Tile

Slices
(also in H.264/AVC)







*D. Finchelstein, V. Sze, A. P. Chandrakasan, "Multi-core Processing and Efficient On-chip Caching for H.264 and Future Video Decoders," *IEEE Trans. CSVT*, 2009

Credit: Vivienne Sze & Madhukar Budagavi, ISCAS 2014 T

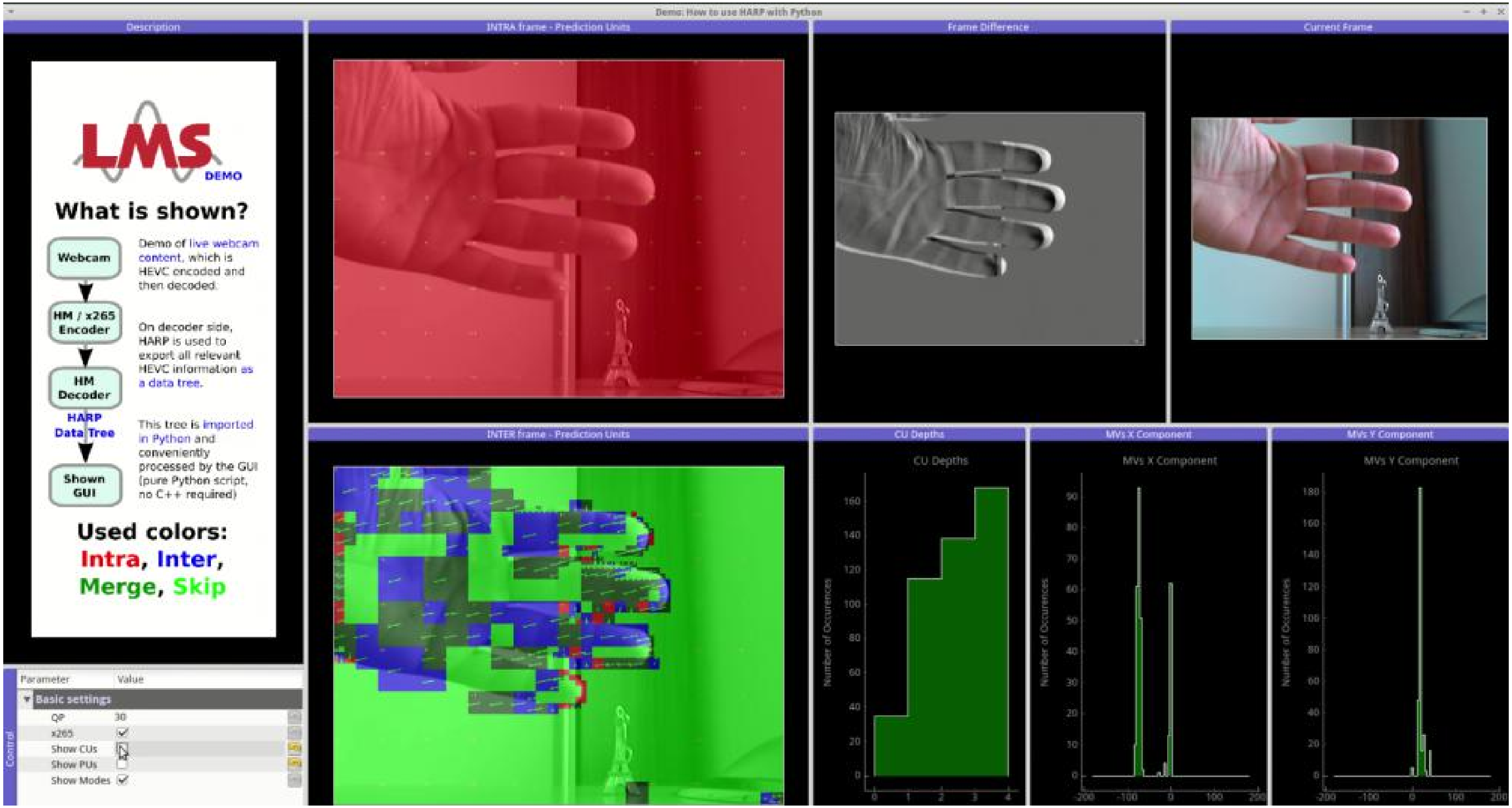
HEVC Resources

- Main Spec:
  - http://www.itu.int/ITU-T/recommendaBons/rec.aspx?rec=11885
- T-CSVT Special Issue:
  - 2012: Combined Issue on HEVC Standard and Research: http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6403920
  - 2016: Special Issue on HEVC Extensions and Efficient HEVC Implementations:

    http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=7372356
- Springer Book:
  - V. Sze, M. Budagavi, G. J. Sullivan (Editors), "High Efficiency Video Coding (HEVC): Algorithms and Architectures," Springer, 2014, http://www.springer.com/engineering/signals/book/978-3-319-06894-7
- HM (open source software):
  - https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSovware/

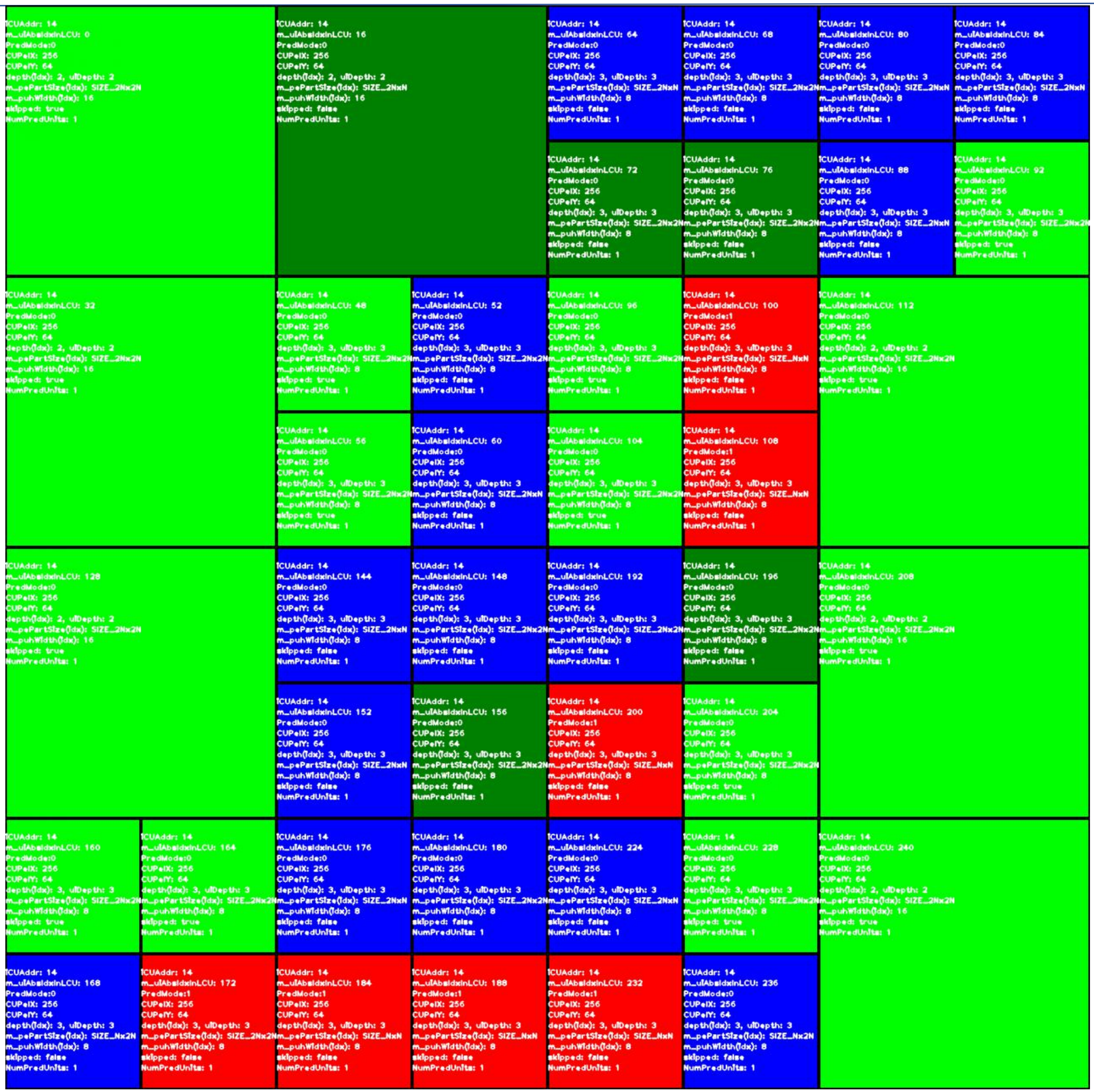
HARP

- HARP: http://www.lms.lnt.de/HARP/
- Visualizing the mode decision and coding results



CTU Close UP

- CTU modes
  - PU modes: Intra, Inter, Merge, SKIP
  - TU modes
  - MVs



Outline

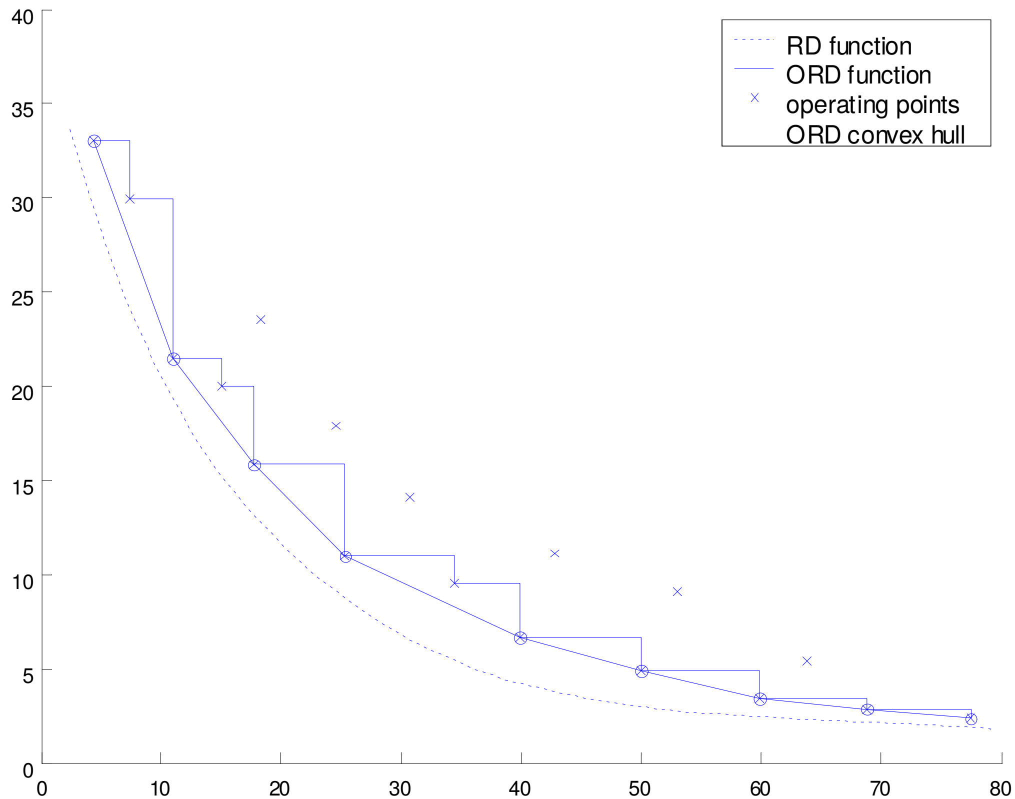
- Recap of HEVC system
- R-D Optimization Basis
- Summary

R-D Optimization in Video Coding

- An image/video coding system needs to decide many different parameters for each frame and each MB:
  - Intra/Inter Prediction modes
  - CTU/MB segmentation
  - Quantization step sizes
  - …
- How to achieve the optimal decision?
- The answer depends on the objective.
  - Minimizing Distortion ?
  - Minimizing Delay ?
  - Minimizing Storage ?

Operational Rate-Distortion Theory

❑ Operational R-D optimization:



- Gives the "operational" R-D performance curve.
- $\{Q_j\}$: operating points associated with coding decisions and parameters
- Optimization: select operating points that minimizes distortion for a given rate, or minimizing rate for a given distortion.

$$R_{op}(D) = \min_{Q_j} R(Q_j),\ s.t.\ D(Q_j) \le D$$

Joint source channel coding optimization

- When transmitted a compressed image/video over network, need to add channel coding
- The joint source channel coding problem:

$$\min_{\substack{\text{source parameters}\\ \text{channel parameters}}} E(D) \quad \text{subject to } R_{source} + R_{channel} \le R_{target}$$

- E(D): expected distortion
- If modem is involved, we also need to optimize the modem parameters (power, modulation constellation etc)

Storage constraint: budget-constrained allocation

- Storage (rate) constraints:

$$\sum_i r_{i,x(i)} \le R_{target}$$

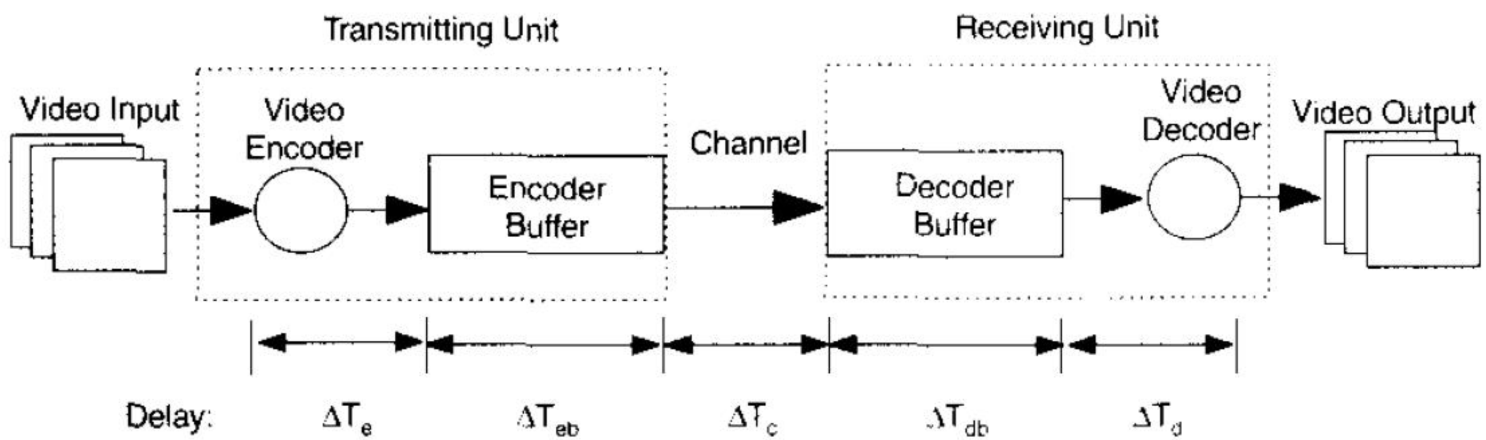
$x(i)$: quantizer index for data unit i.

$r_{i,x(i)}$: rate of unit i with quantizer x(i).

$d_{i,x(i)}$: dist. of unit i with quantizer x(i).

- Examples of objectives:
  - Minimizing MSE (average): $\min \sum_i d_{i,x(i)}$
  - Minimax (max): $\min\left(\max_i d_{i,x(i)}\right)$

Delay-constrained allocation

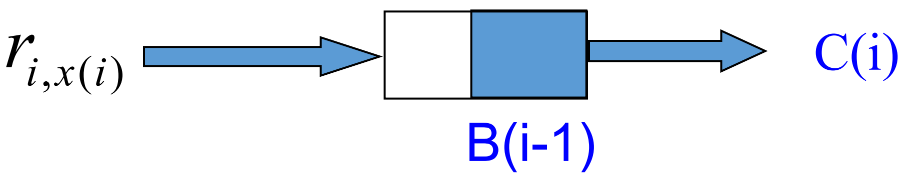


- Find the optimal set of quantizer x(i) such that
  - Each coding unit i encoded at time t_i is received by the decoder before its deadline $t_i + \delta_i$
  - A given distortion metric is minimized.
- For simplicity, assume $\delta_i = \Delta T$
  - If frame rate is f, there will be $\Delta N = f\ \Delta T$ frames at various places in the system.
- The delay constraint can be met by controlling the encoder buffer

Delay-constrained allocation

- C(i): channel rate during the i-th coding interval.
- B(i): encoder buffer level at time i.

$$B(i) = \max\left(B(i-1) + r_{i,x(i)} - C(i),\ \ 0\right).$$



- Transmission of the bits for frame i has to be completed at least within the next ΔN frames to guarantee a max delay of ΔT.
- So B(i) has to satisfy

$$B(i) \le \sum_{k=i+1}^{i+\Delta N} C(k), \quad \text{for all i.}$$

Buffer-constrained allocation

- So the delay-constrained problem becomes a buffer-constrained problem ➔
- Find the optimal quantizer x(i) for each i such that the buffer occupancy

$$B(i) = \max\left(B(i-1) + r_{i,x(i)} - C(i),\ \ 0\right)$$

satisfies the constraint

$$B(i) \le \sum_{k=i+1}^{i+\Delta N} C(k), \quad \text{for all i,}$$

and some distortion metric is minimized.

Multi-User Problem

- Coding for sharing a common communication bottleneck

$$\min_{\{x_k\}} \sum_k D_k(x_k), s.t., \sum_k R(x_k) \le C$$

  - Compute a set of resource allocation $\{x_k\}$ over the scheduling period, for multiple users such that the total/average distortion is minimized
  - And the rate constraint is satisfied
- Leads to a resource pricing solution, that shall be discussed later.
- Lagrangian relaxation, the Lagrangian becomes the resource price.

Lagrangian Method

- To solve the problem of

$$\min \sum_i d_{i,x(i)} \quad \text{subject to} \quad \sum_i r_{i,x(i)} \le R_{target}$$

The Lagrangian method minimizes

$$\min\left(\sum_i d_{i,x(i)} + \lambda \sum_i r_{i,x(i)}\right)$$

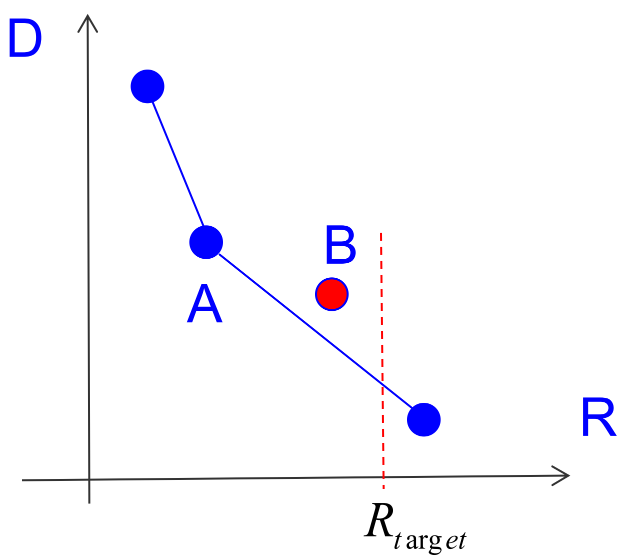
This is equivalent to

$$\sum_i \min\left(d_{i,x(i)} + \lambda r_{i,x(i)}\right)$$

We showed before that this implies that each unit operates at the *same slope* in its R-D curve.

Lagrangian Method

- Problem: The Lagrangian method can only select the operating points that lie on the convex hull, but not above it ➔ the Lagrangian solution is only an approximation of the optimal solution.
- The loss of optimality could be a severe problem if the points on the curve are not dense enough.



Example: for the given target rate, the solution of Lagrangian method is point A, but point B has lower distortion and still meet the rate constraint.

➔ Lagrangian sol is not optimal.

Convex Optimization

- An optimization problem has the general format

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \ldots, m. \end{array}$$

- Convex optimization: the objective and constraint functions are both convex, ie,

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

for all $x, y \in \mathbf{R}^n$ and all $\alpha, \beta \in \mathbf{R}$ with $\alpha + \beta = 1$, $\alpha \geq 0$, $\beta \geq 0$.

- "There is in general no analytical formula for the solution of convex optimization problems, but there are very effective methods for solving them. Interior-point methods work very well in practice."
- If a practical problem as a convex optimization problem, then you have solved the original problem."
- "Recognizing convex optimization problems, or those that can be transformed to convex problems, could be very challenging."

Convex Optimization Course Work

- Very good online course coverage on this topic:
  - Course video by Prof. Stephen Boyd:
  http://web.stanford.edu/class/ee364a/videos.html
- Ref:
  - Slides: http://stanford.edu/~boyd/cvxbook/bv_cvxslides.pdf
  - The book: Stephen Boyd, Convex Optimization, Chapter 5, http://stanford.edu/~boyd/cvxbook/

Convex Optimization
Stephen Boyd and Lieven Vandenberghe

Cambridge University Press

Convex Optimization

- Convex optimization also plays an important role in nonconvex problems.
- Initialization for local optimization:
  - Approximate the problem by a convex problem
  - Use the solution of convex approximation as the starting point for a local optimization method
- Convex heuristics for nonconvex optimization
  - Example: finding a sparse vector that satisfies some constraints
  - This is the case in compressed sensing
- Bounds for nonconvex global optimization
  - Relaxation: each nonconvex constraint is replaced with a looser, but convex, constraint.
  - Lagrangian relaxation: the Lagrangian dual problem is solved, which is convex, and provides a lower bound on the optimal value of the nonconvex problem.

Affine set and convex set

- **Affine set:** A set C is affine if the line **through** any two distinct points in C lies in C, i.e., if for any x1, x2 in C and θ, we have

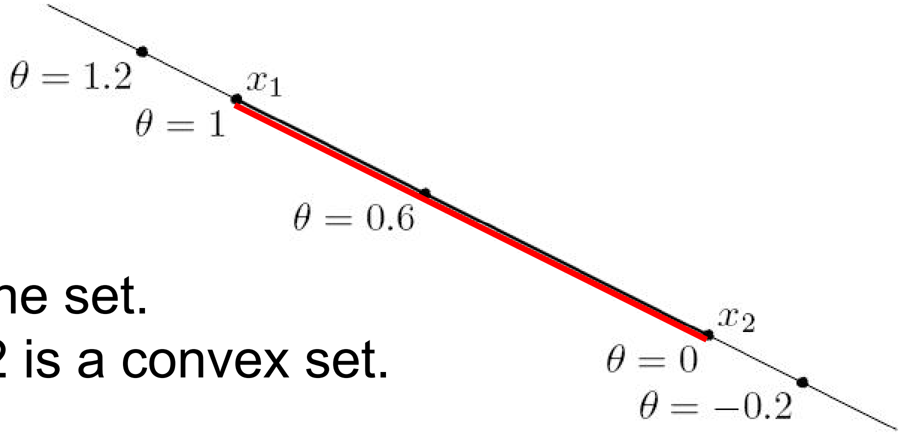
  θ x1 +(1−θ) x2 in C.

- Note: θ can be negative.

- Convex set: A set C is convex if the line segment **between** any two points in C lies in C, i.e., if for any x1, x2 in C and any θ with $0 \le \theta \le 1$, we have

$$\theta\, x1 + (1-\theta)x2 \ \text{ in } \ C.$$

Note: θ is non-negative!



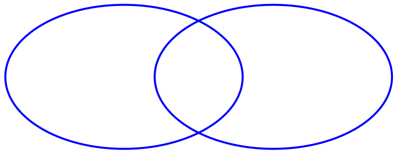
Example:

The line through x1, x2 is an affine set.

The segment between x1 and x2 is a convex set.

Operations that preserve convexity

- **Intersection**: If S1 and S2 are convex, then S1 ∩ S2 is convex.



- **Affine function**: a function f is affine if it is a sum of a linear function and a constant, i.e., $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$.
- Proof:

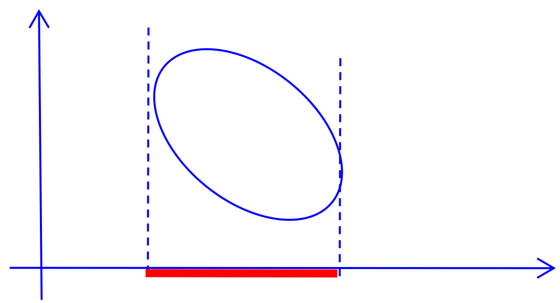
$$\begin{aligned}
&\theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2) \\
&= \theta(\mathbf{A}\mathbf{x}_1 + \mathbf{b}) + (1-\theta)(\mathbf{A}\mathbf{x}_2 + \mathbf{b}) \\
&= \mathbf{A}(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2) + \mathbf{b}.
\end{aligned}$$

- Suppose S is a convex set and f is an affine function. Then f(S) = {f(x) | x in S} is convex (a convex set is still convex after scaling and transition)

Operations that preserve convexity

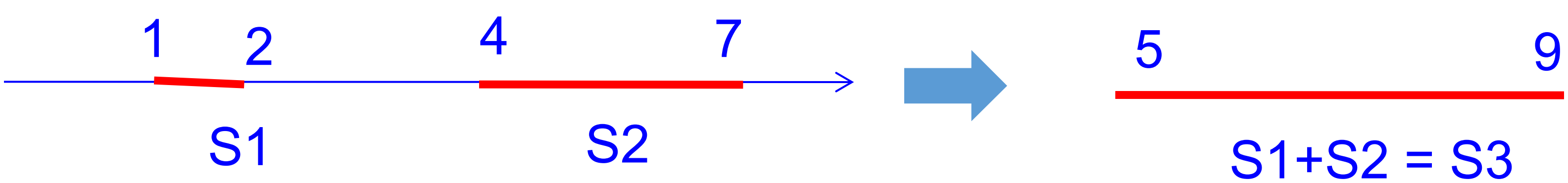
❑Examples:

- The projection of a convex set onto some of its coordinates is convex



- The sum of two convex sets is convex, where the sum is defined as

$$S_1 + S_2 = \{x + y \mid x \in S_1,\ y \in S_2\}.$$



Lagrangian

- Consider a standard optimization problem (not necessarily convex)

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0, & i = 1, \ldots, m \\ & h_i(x) = 0, & i = 1, \ldots, p, \end{array}$$

- Assume the optimal value of $f_0(x)$ is p*.
- The Lagrangian is defined as

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

- Group all λ_i into a vector λ, and all ν_i into vector v.
- v and λ are called Lagrangian multipliers or dual variables.
- Note: The Lagrangian is unconstrained.

The Lagrangian dual function

- The Lagrange dual function (or just dual function) g is the minimum value of the Lagrangian over x:

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right).$$

- Important: The **dual function is always concave**, even if the original problem (the primal problem) is not convex.
- This is one benefit of working with the dual function, especially when the original problem is difficult to solve.

- Notation: "*inf*" stands for infimum. The infimum of a subset is the greatest element, not necessarily in the subset, that is <= all elements of the subset. *Inf* is also called greatest lower bound (GLB).
- Example:

$$\inf \{x \in \mathbb{R} : 0 < x < 1\} = 0.$$

Duality Gap

- Duality problem solves the primal problem when duality gap is zero.
  - Feasible region: (*u*, *t*) in *F*

$$F:\{(u,t)\,|\,u=f_1(x),t=f_0(x),\forall x\in DOM\ f_0\}$$

- Dual problem:

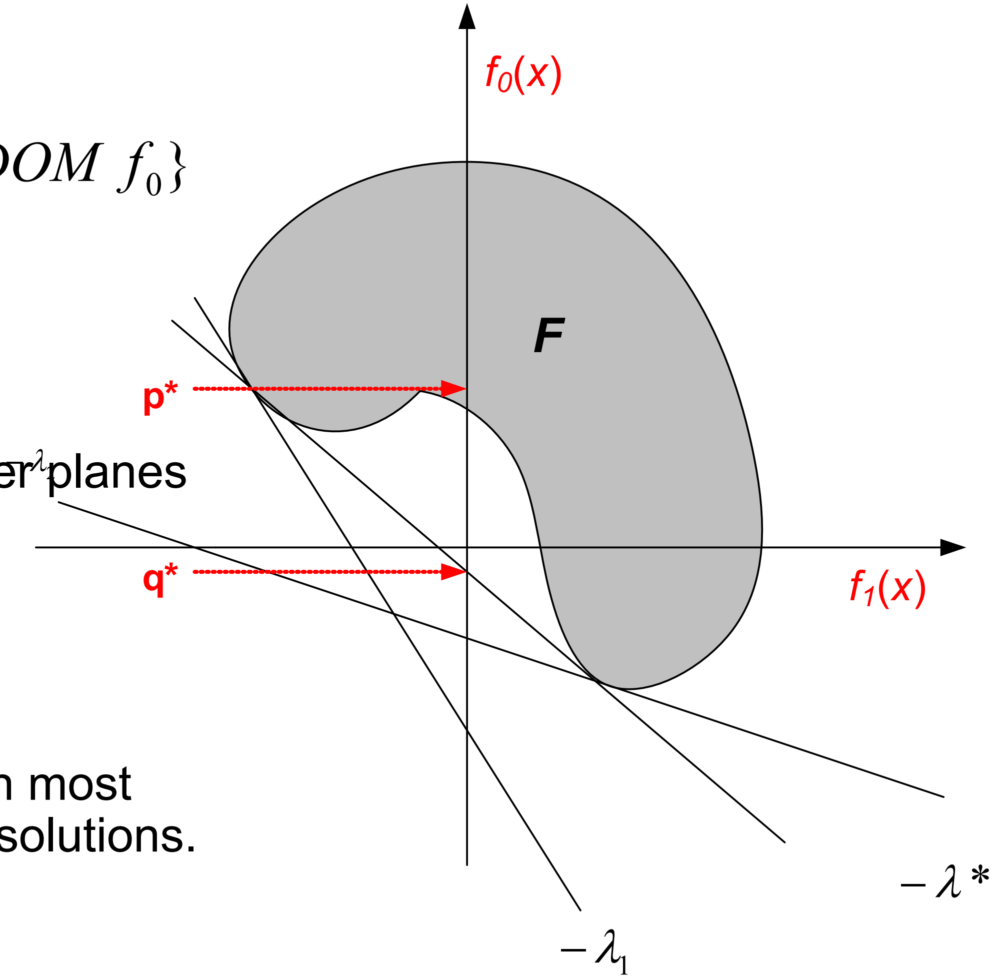
$$\max_{\lambda\geq 0} g(\lambda)=\max_{\lambda\geq 0}\{\inf_{u,t\in F}(t+\lambda u)\}$$

- Searching on a set of supporting hyper planes with slope - λ

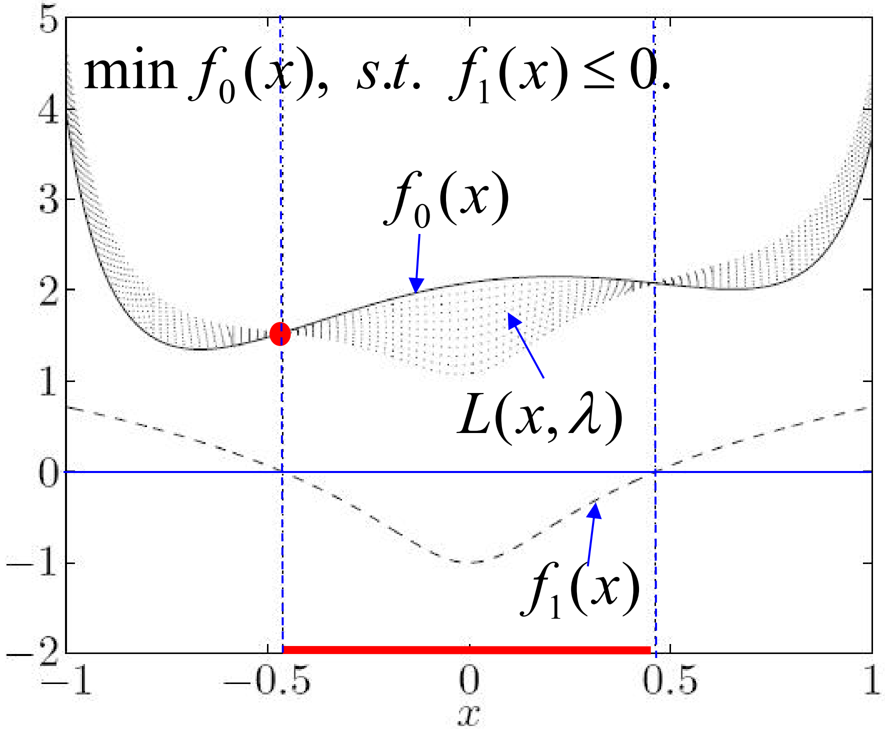
- Duality gap: *p**- *q**

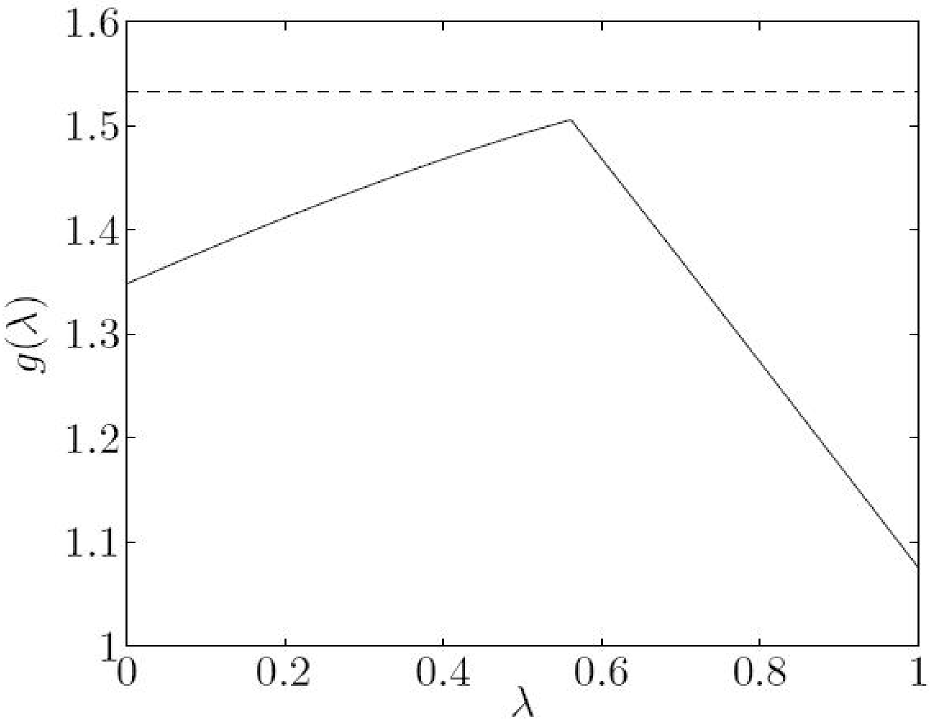
- “Convex Hull” approximation typical in most Lagrangian relaxed R-D Optimization solutions. Penalty is this duality gap

$$p^*=\min_{x\in C} f_0(x),\ s.t.\ f_1(x)\leq 0$$



Lower bound on optimal value





Solid line: $f_0(x)$
Dashed line: f1(x)
Feasible set: [-0.46, 0.46] (f1(x) <=0)
Optimal solution: x* = -0.46, p* = 1.54
Dotted line: L(x, λ) for λ = 0.1, 0.2, …, 1.0.
The minimum of each dotted line is <= p*.

Dual function g(λ)= inf L(x, λ).
(Note that fi(x) is not convex,
But g(λ) is concave!)
Dashed line: p*.

Example

$$\begin{array}{ll} \text{minimize} & x^T x \\ \text{subject to} & Ax = b, \end{array}$$

- Using the Lagrangian multiplier method, we define

$$L(x,\nu) = x^T x + \nu^T (Ax - b),$$

- Lagrangian is unconstrained ➔ the optimal sol. should be a stationary point (there is no boundary point)

$$\nabla_x L(x,\nu) = 2x + A^T \nu = 0, \quad \Rightarrow \quad x = -(1/2) A^T \nu.$$

- Plugging into the constraint Ax = b ➔

$$\mathbf{v} = -2(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}$$

- So the optimal x is $\mathbf{x} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}$
- The min obj is $\mathbf{b}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}$

Lagrangian dual problem

- The dual function call tell how suboptimal a given feasible point is, without knowing the exact value of p*.
- A natural question: What's the best lower bound that can be obtained from the Lagrangian dual function?
- This is answered by the Lagrangian dual problem:

$$\boxed{\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0. \end{array}}$$

- The dual problem is a convex problem, since the objective function is concave, and the constraint λ >= 0 is convex set, even if the primal problem is not convex.
- Another benefit: dual problem usually has lower dimension than primal problem.
- If λ* and v* are optimal, they are called dual optimal.

Lagrangian dual problem

- Example:

$$\begin{array}{ll} \text{minimize} & x^T x \\ \text{subject to} & Ax = b, \end{array}$$

- We know that

$$g(\nu) = -(1/4)\nu^T A A^T \nu - b^T \nu,$$

- The dual problem is max (g(v)), which is a unconstrained problem (easier)
- To find the max value:

$$dg(\mathbf{v})/d\mathbf{v} = -1/2(\mathbf{AA}^T)^{-1}\mathbf{v} - \mathbf{b} = 0$$

$$\Rightarrow \quad \mathbf{v} = -2(\mathbf{AA}^T)^{-1}\mathbf{b}$$

The max g(v) is: $\mathbf{b}^T(\mathbf{AA}^T)^{-1}\mathbf{b}$

same as the min of the primal problem.

Weak Duality

- Let the optimal value of the Lagrange dual problem be d*.
- This is the best lower bound on p*.
- By definition of dual function, we have

$$d^* \leq p^*.$$

This is called weak duality.

- Optimal duality gap: p* - d*.
- Duality gap provides a good termination criterion if iteration is used.

$$f_0(x^{(k)}) - g(\lambda^{(k)}, \nu^{(k)}) \leq \epsilon_{\text{abs}}$$

Strong Duality

- If d* = p*, then we say that strong duality holds.
- This is usually (not always) true if the primal problem is convex.
- In this case, the dual problem yields the same solution as primal problem, but is usually easier to solve.

- This is the case in the previous example:

$$\begin{array}{ll} \text{minimize} & x^T x \\ \text{subject to} & Ax = b, \end{array}$$

Complementary Slackness

- Suppose strong duality holds. Let x* be a primal optimal and (λ*, ν*) be dual optimal ➔

$$
\begin{aligned}
f_0(x^\star) &= g(\lambda^\star, \nu^\star) \\
&= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^\star f_i(x) + \sum_{i=1}^{p} \nu_i^\star h_i(x) \right) \\
&\leq f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star h_i(x^\star) \\
&\leq f_0(x^\star).
\end{aligned}
$$

The two inequalities in this chain holds with equality. What does this mean ?

$$
\sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) = 0.
$$

Complementary slackness

$$\sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) = 0.$$

λi* >=0 and fi(x) <=0 ➔ $\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m.$

This is called complementary slackness.

This implies that

$$\boxed{\begin{aligned} \lambda_i^\star > 0 &\implies f_i(x^\star) = 0, \\ f_i(x^\star) < 0 &\implies \lambda_i^\star = 0. \end{aligned}}$$

Otherwise the strong duality (d* = p*) is not true.

Karush–Kuhn–Tucker (KKT) Condition

- ❑Let x* be a primal optimal and (λ*, ν*) be dual optimal with zero duality gap, ie, strong duality holds:
- ❑x* minimizes the unconstrained L(x, λ*, ν*) over x
- ❑➔ its gradient must vanish at x* (there is no boundary point)

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

KKT Condition

- Thus x* satisfies the following necessary conditions:

$$
\begin{aligned}
f_i(x^\star) &\le 0, \quad i = 1, \ldots, m \\
h_i(x^\star) &= 0, \quad i = 1, \ldots, p \\
\lambda_i^\star &\ge 0, \quad i = 1, \ldots, m \\
\lambda_i^\star f_i(x^\star) &= 0, \quad i = 1, \ldots, m \\
\nabla f_0(x^\star) + \textstyle\sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^p \nu_i^\star \nabla h_i(x^\star) &= 0,
\end{aligned}
$$

- This is called Karush-Kuhn-Tucker (KKT) conditions:
  - Conditions 1-2: Primal feasibility
  - Condition 3: Dual feasibility
  - Condition 4: Complementary slackness - when constraint is tight, lambda=0
  - Condition 5: Stationarity

- Significance: It is a generalization of the Lagrange multiplier method to inequality constraints.
  - Can be used to find the solution of the problem.

W. Karush (1939). "Minima of Functions of Several Variables with Inequalities as Side Constraints". M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois.

KKT Condition

- In general, KKT condition is not sufficient for optimality, and addtional information is necessary, such as the Second Order Sufficient Conditions.

- However, when the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal.

- The KKT conditions play an important role in optimization. In some cases it is possible to solve the KKT conditions (and therefore the optimization problem) analytically.

- More generally, many algorithms for convex optimization are conceived as, or can be interpreted as, methods for solving the KKT conditions.

KKT Condition

- **Example:**

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P x + q^T x + r \\ \text{subject to} & Ax = b, \end{array}$$

- Solution: There is no inequality constraint, so only two KKT equations (ie, Lagrangian multipler mtd)

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

$$\rightarrow \quad Ax^\star = b, \qquad Px^\star + q + A^T \nu^\star = 0,$$

$$\rightarrow \quad \begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^\star \\ \nu^\star \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}.$$

The solution can be obtained by solving the equations.

Summary

- RD Optimization
  - Lagrangian Operational RD Optimization: can only trace out the convex hull of the R-D plane.
  - Convert the original constrained problem into un-constrained Dual problem, searching the Lagrangian for convex-hull optimal solution
- Next Class:
  - Will cover DP, and show a practical problem solved by Lagrangian Relaxation as outer loop of control, and DP as inner loop of control.