

Optimization: Convexity, Lagrange Multipliers, KKT, and Duality

Lecture Notes Consolidated with Claude AI — CS6140 Machine Learning

Based on V. Pavlu’s notes and slides for this course (*Convex optimization, Constrained convex optimization*), Boyd & Vandenberghe, *Convex Optimization* (Cambridge, 2004; Chapters 2–5, 9–11), and the CS229 *Convex Optimization Overview* notes (Stanford).

Notation. The optimization variable is $\mathbf{w} \in \mathbb{R}^D$, a $D \times 1$ column vector (as for the weight vector in the other notes); a datapoint $\mathbf{x}$ is a $1 \times D$ row vector, so $\mathbf{x}\mathbf{w}$ is a number. The gradient $\nabla f(\mathbf{w})$ is a $D \times 1$ column and the Hessian $\nabla^2 f(\mathbf{w})$ is $D \times D$. Inequality constraints are $g_i(\mathbf{w}) \leq 0$, $i = 1..k$, with multipliers $\alpha_i \geq 0$; equality constraints are $h_j(\mathbf{w}) = 0$, $j = 1..m$, with multipliers β_j . A tilde marks the solution, $\tilde{\mathbf{w}}$; p^* and d^* are the optimal primal and dual values.

1 Why Optimization

Almost every model in this course is trained by *minimizing* something, sometimes subject to constraints:

- least squares: $\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2$; logistic regression: $\min_{\mathbf{w}} \sum_i \log(1 + e^{-y_i \mathbf{x}_i \mathbf{w}})$ (Regression and GD/Logistic Regression notes);
- PCA: $\max_{\mathbf{e}} \mathbf{e}^T \Sigma \mathbf{e}$ subject to $\mathbf{e}^T \mathbf{e} = 1$ (PCA note): an *equality* constraint;
- SVM: $\min \frac{1}{2} \|\mathbf{w}\|^2$ subject to $y_i(\mathbf{x}_i \mathbf{w} + b) \geq 1$ (SVM note): many *inequality* constraints;
- lasso: minimize the loss with $\|\mathbf{w}\|_1$ small — a constraint in disguise (§9).

This note collects the tools behind all of them: what makes a problem easy (*convexity*, §2), how to solve it without constraints (§3), how constraints enter through *Lagrange multipliers* (§4) and the *KKT conditions* (§5), and the *dual problem* (§6). Section 7 has the full proofs and can be skipped on first reading.

2 Convex Sets and Convex Functions

2.1 Convex Sets

A set $X \subseteq \mathbb{R}^D$ is **convex** if the segment between any two of its points stays inside it:

$$\mathbf{w}_1, \mathbf{w}_2 \in X, \lambda \in [0, 1] \Rightarrow \lambda \mathbf{w}_1 + (1 - \lambda) \mathbf{w}_2 \in X. \quad (1)$$

Examples: a half-space $\{\mathbf{w} : \mathbf{a}\mathbf{w} \leq b\}$; a ball $\{\|\mathbf{w}\| \leq r\}$; any **intersection** of convex sets (so a polyhedron, Figure 1 right). In particular the feasible set $\{\mathbf{w} : g_i(\mathbf{w}) \leq 0 \forall i\}$ is convex whenever every g_i is a convex function, because each $\{g_i \leq 0\}$ is convex (a “sublevel set”).

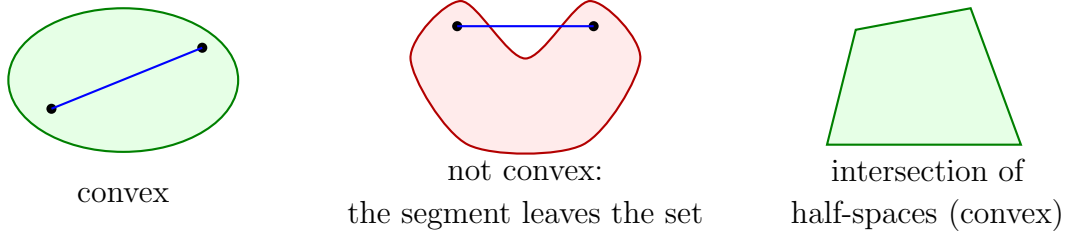


Figure 1: Convex and non-convex sets.

2.2 Convex Functions

A function f is **convex** on a convex set X if for all $\mathbf{w}_1, \mathbf{w}_2 \in X$ and $\lambda \in [0, 1]$

$$f(\lambda \mathbf{w}_1 + (1 - \lambda) \mathbf{w}_2) \leq \lambda f(\mathbf{w}_1) + (1 - \lambda) f(\mathbf{w}_2) : \quad (2)$$

the chord between two points of the graph lies above the graph (Figure 2a). It is **strictly convex** if the inequality is strict whenever $\mathbf{w}_1 \neq \mathbf{w}_2$ and $0 < \lambda < 1$; **concave** if $-f$ is convex. Applying (2) repeatedly gives **Jensen's inequality**: for weights $\lambda_i \geq 0$ with $\sum_i \lambda_i = 1$ (a distribution),

$$f\left(\sum_i \lambda_i \mathbf{w}_i\right) \leq \sum_i \lambda_i f(\mathbf{w}_i). \quad (3)$$

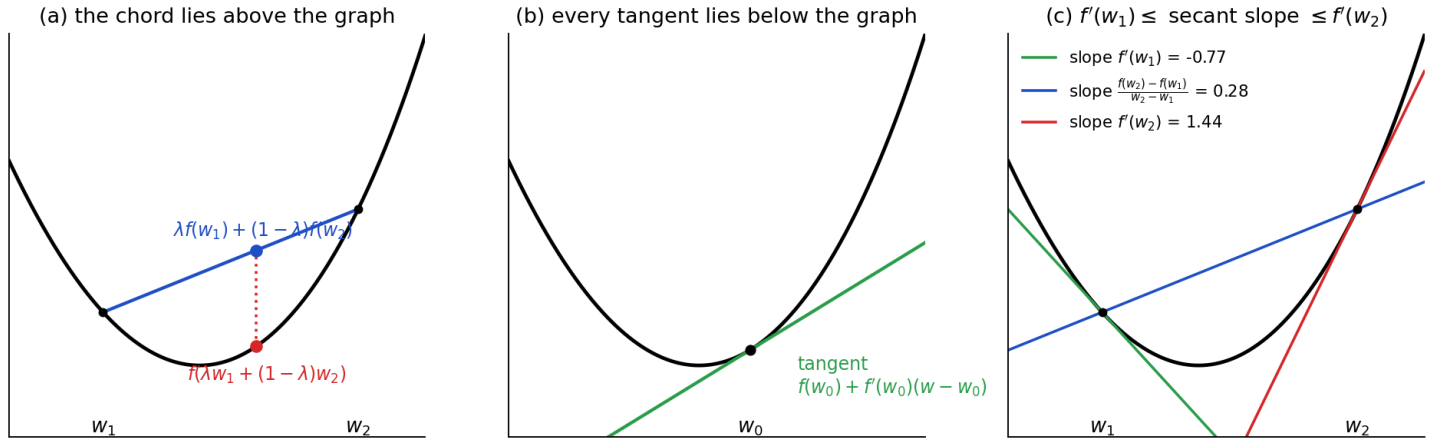


Figure 2: A convex function. (a) Definition (2). (b) First-order condition (4). (c) The slopes inequality (5): the tangent slope at w_1 , the chord slope, and the tangent slope at w_2 increase in that order.

2.3 Derivatives of Convex Functions

If f is convex and differentiable, **every tangent lies below the graph** (Figure 2b):

$$f(\mathbf{w}) \geq f(\mathbf{w}_0) + \nabla f(\mathbf{w}_0)^T (\mathbf{w} - \mathbf{w}_0) \quad \text{for all } \mathbf{w}, \mathbf{w}_0. \quad (4)$$

This is the most useful fact in the note: a local, first-order quantity (the gradient at $\mathbf{w}_0$) gives a *global* lower bound on f . In one variable, applying it at w_1 and at w_2 gives the **slopes inequality** from the slides (Figure 2c): for $w_1 < w_2$,

$$f'(w_1) \leq \frac{f(w_2) - f(w_1)}{w_2 - w_1} \leq f'(w_2), \quad (5)$$

so f' is nondecreasing (strictly increasing if f is strictly convex). By the **mean value theorem** the chord slope is attained in between: there is $w = \lambda w_1 + (1 - \lambda)w_2$, $\lambda \in [0, 1]$, with $f(w_2) - f(w_1) = f'(w)(w_2 - w_1)$. *Proof of (4) in one variable.* By (2) with $w_0 + \lambda(w - w_0) = (1 - \lambda)w_0 + \lambda w$: $f(w_0 + \lambda(w - w_0)) \leq f(w_0) + \lambda(f(w) - f(w_0))$. Subtract $f(w_0)$, divide by $\lambda > 0$, and let $\lambda \rightarrow 0$: the left side tends to $f'(w_0)(w - w_0)$, so $f'(w_0)(w - w_0) \leq f(w) - f(w_0)$. In D dimensions apply this to $\phi(s) = f(\mathbf{w}_0 + s(\mathbf{w} - \mathbf{w}_0))$. $\square$

Second-order condition. A twice-differentiable f is convex if and only if $f'' \geq 0$ (one variable), or the Hessian is positive semidefinite, $\mathbf{v}^T \nabla^2 f(\mathbf{w}) \mathbf{v} \geq 0$ for all $\mathbf{v}$ and $\mathbf{w}$ (several variables). Example: $\|X\mathbf{w} - \mathbf{y}\|^2$ has Hessian $2X^T X$, and $\mathbf{v}^T X^T X \mathbf{v} = \|X\mathbf{v}\|^2 \geq 0$ — the same argument that made XX^T a valid kernel matrix (Kernels note). Strict convexity does not require $f'' > 0$: w^4 is strictly convex with $f''(0) = 0$.

2.4 Convex Functions in This Course

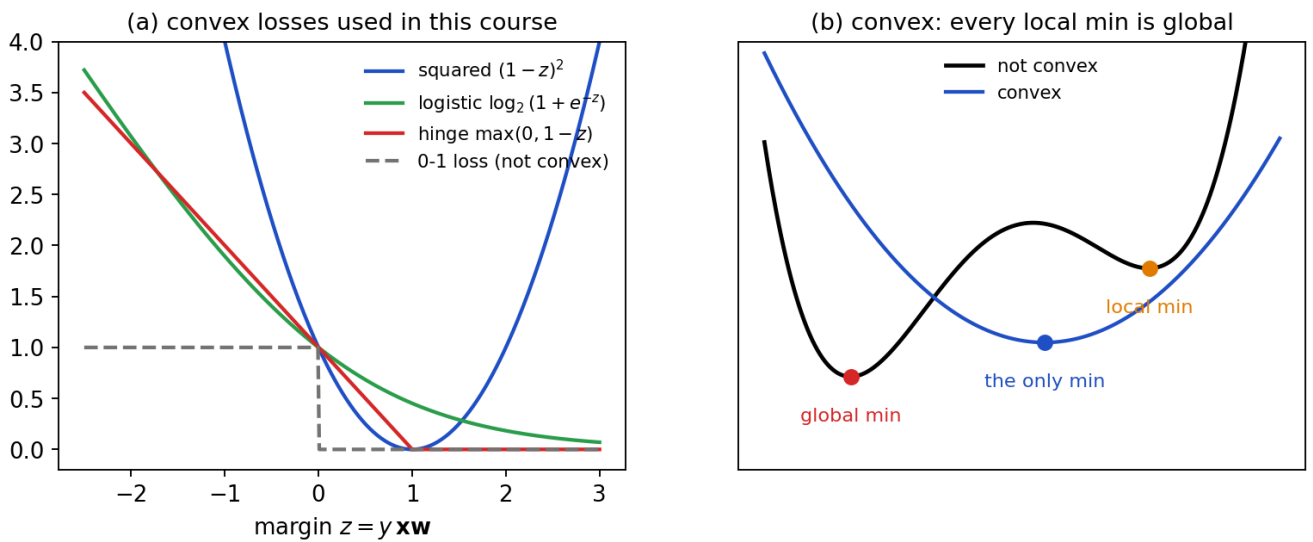


Figure 3: (a) The squared, logistic and hinge losses are convex in the margin z , and hence in $\mathbf{w}$; the 0-1 loss is not. (b) A non-convex function can trap a descent method in a local minimum; a convex one has only global minima.

Most losses we minimize are convex in $\mathbf{w}$ (Figure 3a): $\|\mathbf{w}\|^2$, $\|\mathbf{w}\|_1$, squared loss, logistic loss, and the hinge loss $\max(0, 1 - y\mathbf{x}\mathbf{w})$ (convex but not differentiable at the kink). Rather than checking (2) each time, use the rules that **preserve convexity**:

- nonnegative weighted sums $\sum_i c_i f_i$, $c_i \geq 0$ (a loss summed over datapoints, plus a regularizer);
- composition with an affine map, $f(\mathbf{x}\mathbf{w} + b)$ (each loss is a convex function of the margin, which is linear in $\mathbf{w}$);
- pointwise maximum $\max(f_1, f_2)$ (hinge = $\max(0, 1 - z)$; $|w| = \max(w, -w)$).

2.5 Why Convexity Matters

Theorem (local = global). *If f is convex on a convex set X , every local minimum of f on X is a global minimum. If f is also differentiable and $X = \mathbb{R}^D$, then $\nabla f(\tilde{\mathbf{w}}) = 0$ if and only if $\tilde{\mathbf{w}}$ is a global minimum.*

Proof. Let $\tilde{\mathbf{w}}$ be a local minimum and suppose some $\mathbf{w} \in X$ has $f(\mathbf{w}) < f(\tilde{\mathbf{w}})$. For every $\lambda \in (0, 1]$ the point $\mathbf{w}_\lambda = (1 - \lambda)\tilde{\mathbf{w}} + \lambda\mathbf{w}$ is in X and $f(\mathbf{w}_\lambda) \leq (1 - \lambda)f(\tilde{\mathbf{w}}) + \lambda f(\mathbf{w}) < f(\tilde{\mathbf{w}})$. As $\lambda \rightarrow 0$ these points come arbitrarily close to $\tilde{\mathbf{w}}$, contradicting local minimality. For the second claim, (4) at $\mathbf{w}_0 = \tilde{\mathbf{w}}$ with $\nabla f(\tilde{\mathbf{w}}) = 0$ gives $f(\mathbf{w}) \geq f(\tilde{\mathbf{w}})$ for all $\mathbf{w}$; the converse is the usual necessary condition. $\square$

Two more facts from the slides, both one-line consequences of (2):

- **The set of solutions is convex.** If $\mathbf{w}_1, \mathbf{w}_2$ both attain the minimum p^* over a convex feasible set, every point between them is feasible and has $f \leq \lambda p^* + (1 - \lambda)p^* = p^*$, so it is also a solution.
- **Strictly convex $\Rightarrow$ at most one minimizer.** Two different minimizers would have a midpoint with $f < p^*$.

Key point: For a convex problem, “the gradient is zero” (or, with constraints, “the KKT conditions hold”, §5) is not just a hint: it certifies the global optimum. That is why we prefer convex losses, and why logistic regression and the SVM have no bad local minima, while neural networks (later in the course) do.

3 Unconstrained Optimization

3.1 One Variable: Interval Cutting and Newton’s Method

Running example: $f(w) = e^w - 2w$, with $f'(w) = e^w - 2$, $f''(w) = e^w > 0$ (strictly convex), minimum at $w = \ln 2 \approx 0.6931$.

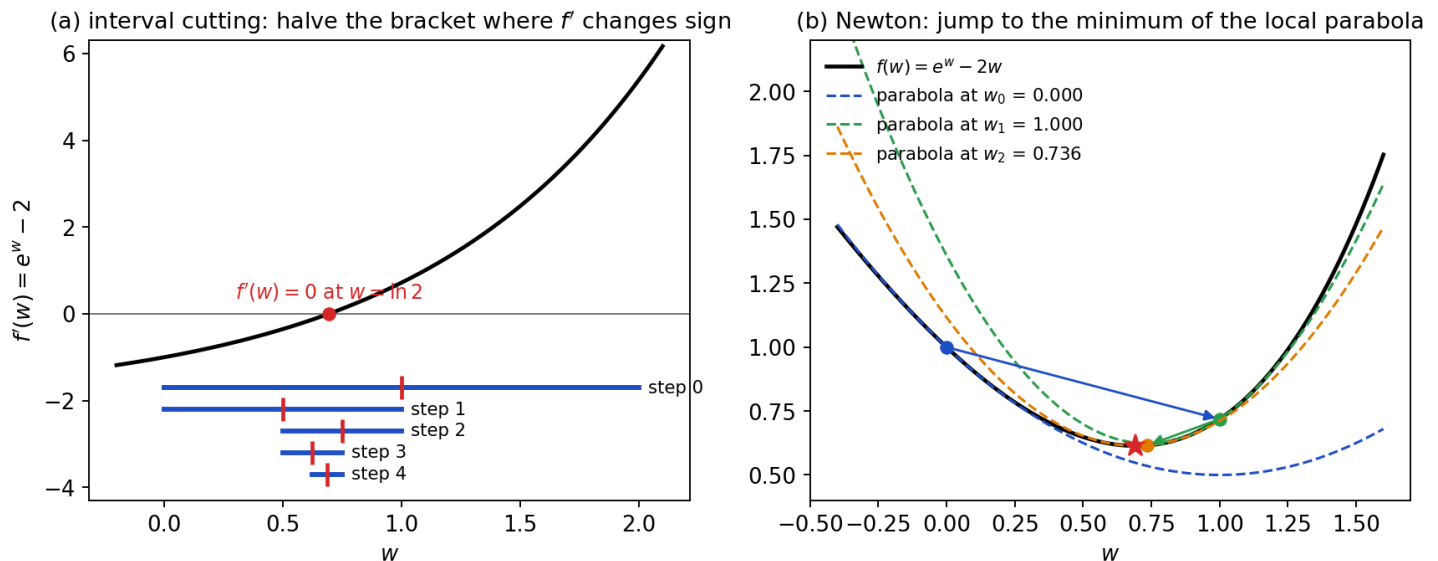


Figure 4: (a) Interval cutting on f' : the bracket $[0, 2]$ is halved five times. (b) Newton’s method from $w_0 = 0$: each step jumps to the minimum of the parabola that matches f, f', f'' at the current point.

- **Interval cutting (bisection).** The minimum is where f' changes sign. Start with $[a, b]$ such that $f'(a) < 0 < f'(b)$; evaluate f' at the midpoint and keep the half where the sign still changes (Figure 4a). The error halves every step: $[0, 2] \rightarrow [0, 1] \rightarrow [0.5, 1] \rightarrow [0.5, 0.75] \rightarrow [0.625, 0.75] \rightarrow [0.6875, 0.75]$.

- **Newton’s method.** Approximate f near w_k by its second-order Taylor parabola $f(w_k) + f'(w_k)(w - w_k) + \frac{1}{2}f''(w_k)(w - w_k)^2$ and jump to its minimum:

$$w_{k+1} = w_k - \frac{f'(w_k)}{f''(w_k)}. \quad (6)$$

From $w_0 = 0$ the iterates are $w_1 = 1$, $w_2 = 0.7358$, $w_3 = 0.69404$, $w_4 = 0.6931476$; their errors are 0.69, 0.31, 0.043, 9×10^{-4} , 4×10^{-7} . Near the minimum the number of correct digits roughly *doubles* every step (**quadratic convergence**), against one extra binary digit per step for bisection.

3.2 Several Variables

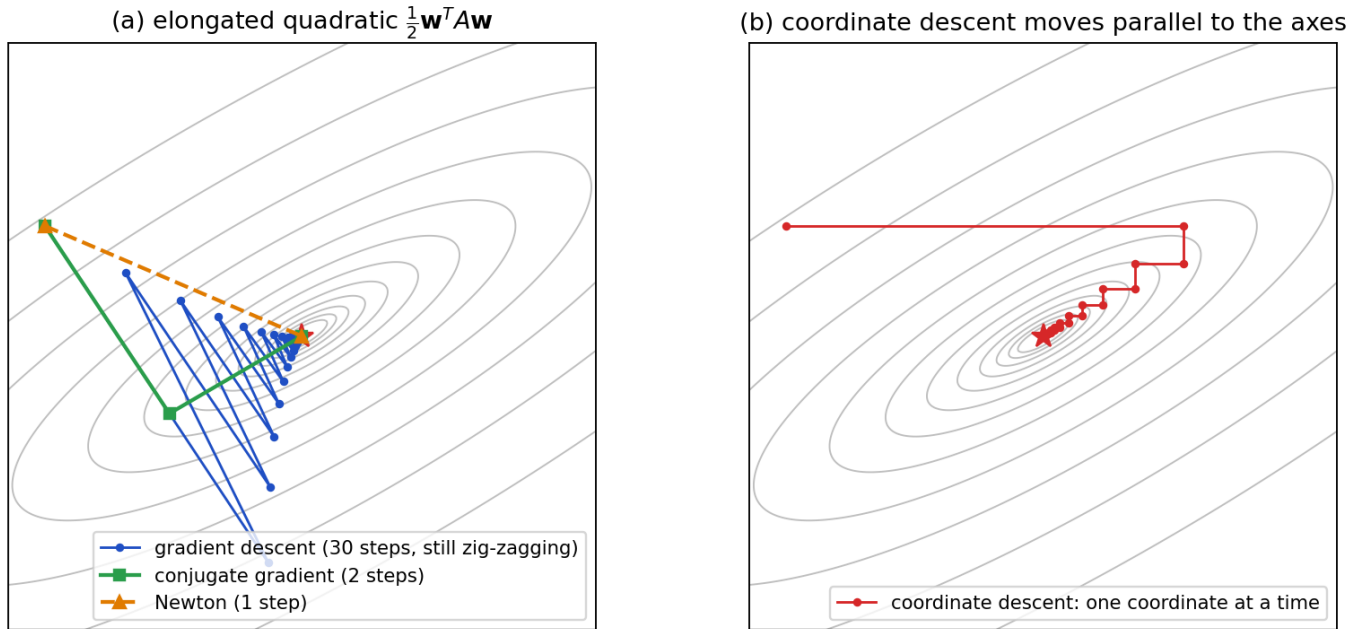


Figure 5: Minimizing an elongated quadratic $\frac{1}{2}\mathbf{w}^T A \mathbf{w}$ (eigenvalues 1 and 12) from the same start. (a) Gradient descent with a fixed step zig-zags across the valley (after 30 steps it is still 0.014 from the minimum); conjugate gradient reaches the minimum exactly in $D = 2$ steps; Newton in one. (b) Coordinate descent minimizes along one axis at a time.

- **Gradient descent** (GD/Logistic Regression note): $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla f(\mathbf{w})$. Cheap per step, but on elongated valleys the gradient points across the valley, not along it, so it zig-zags (Figure 5a).
- **Newton’s method:** $\mathbf{w} \leftarrow \mathbf{w} - (\nabla^2 f(\mathbf{w}))^{-1} \nabla f(\mathbf{w})$. The Hessian rescales the directions, so a quadratic is solved in one step; the cost is forming and solving a $D \times D$ system per step.
- **Conjugate gradient:** on a quadratic $\frac{1}{2}\mathbf{w}^T A \mathbf{w} - \mathbf{b}^T \mathbf{w}$, choose each new direction “ A -conjugate” to the previous ones ($\mathbf{d}_i^T A \mathbf{d}_j = 0$) and minimize exactly along it; this reaches the minimum in at most D steps, using only matrix–vector products.
- **Coordinate descent:** minimize over one coordinate (or a small block) at a time, keeping the others fixed (Figure 5b). Each step is often solvable in closed form. SMO, the SVM solver, is a coordinate method that updates two variables at a time.

4 Equality Constraints: Lagrange Multipliers

4.1 Mary, Cal, and the River

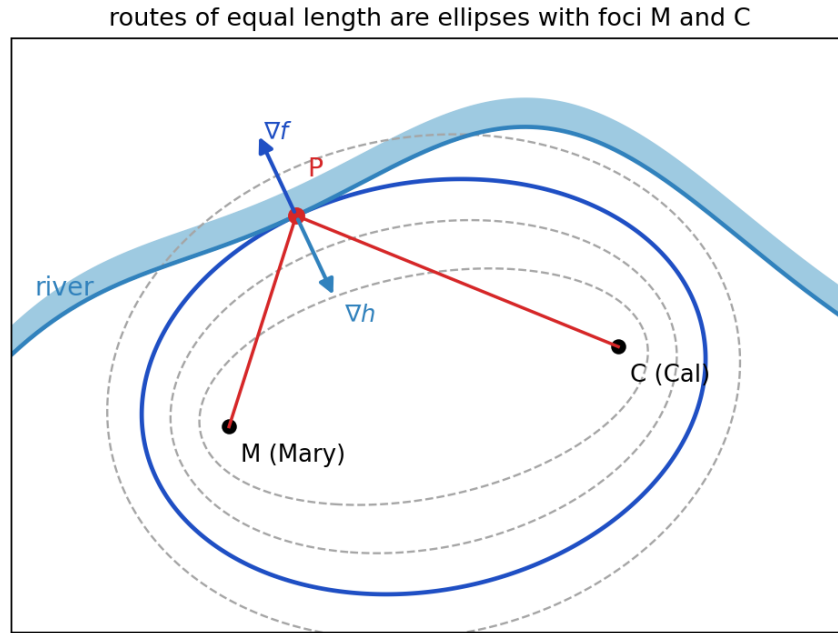


Figure 6: Mary (M) must reach Cal (C) as fast as possible but has to stop at the river first. Routes of equal length $|MP| + |PC|$ are ellipses with foci M and C . The best stop P is where the smallest ellipse that still reaches the river touches it: there the ellipse and the river are tangent, so ∇f and ∇h are parallel (computed: the angle between them is 0).

Mary has a date with Cal; she wants to get there as soon as possible, but she has to stop by the river first. The objective is the route length $f(P) = |MP| + |PC|$ and the constraint is “ P is on the river”, $h(P) = 0$. All stops of the same route length lie on an ellipse with foci M and C . Growing the ellipse from small to large, the first one that touches the river gives the best stop P (Figure 6). At P the ellipse and the river are **tangent**; the gradient of each is perpendicular to its own curve, so the two gradients are **parallel**:

$$\nabla f(P) = -\beta \nabla h(P) \quad \text{for some number } \beta. \quad (7)$$

The proportionality constant β is the **Lagrange multiplier**. (Bonus: $\nabla f(P)$ is the sum of the unit vectors from M and from C to P , so this says the path reflects off the river with equal angles, Heron’s law of reflection.)

4.2 The Lagrangian and Lagrange’s Theorem

For the problem

$$\min_{\mathbf{w}} f(\mathbf{w}) \quad \text{subject to} \quad h_j(\mathbf{w}) = 0, \quad j = 1..m, \quad (8)$$

assign one multiplier β_j to each constraint and define the **Lagrangian**

$$L(\mathbf{w}, \beta) = f(\mathbf{w}) + \sum_j \beta_j h_j(\mathbf{w}). \quad (9)$$

On the feasible set $L = f$. Setting its derivatives to zero reproduces both the tangency condition and the constraints:

$$\nabla_{\mathbf{w}} L(\tilde{\mathbf{w}}, \tilde{\beta}) = \nabla f(\tilde{\mathbf{w}}) + \sum_j \tilde{\beta}_j \nabla h_j(\tilde{\mathbf{w}}) = 0, \quad \frac{\partial L}{\partial \beta_j}(\tilde{\mathbf{w}}, \tilde{\beta}) = h_j(\tilde{\mathbf{w}}) = 0. \quad (10)$$

Theorem (Lagrange). *If $\tilde{\mathbf{w}}$ is a local minimum and the gradients $\nabla h_j(\tilde{\mathbf{w}})$ are linearly independent (a “regular point”), then there exist $\tilde{\beta}_j$ satisfying (10). If moreover f is convex and the h_j are affine, (10) is also sufficient: any $(\tilde{\mathbf{w}}, \tilde{\beta})$ satisfying it gives a global minimum.*

Proof sketch (one constraint). Let $\mathbf{d}$ be any direction tangent to the constraint surface at $\tilde{\mathbf{w}}$, i.e. $\nabla h(\tilde{\mathbf{w}})^T \mathbf{d} = 0$. Since $\nabla h(\tilde{\mathbf{w}}) \neq 0$, there is a curve $\mathbf{c}(s)$ inside $\{h = 0\}$ with $\mathbf{c}(0) = \tilde{\mathbf{w}}$ and $\mathbf{c}'(0) = \mathbf{d}$ (implicit function theorem). The function $s \mapsto f(\mathbf{c}(s))$ has a local minimum at $s = 0$, so its derivative $\nabla f(\tilde{\mathbf{w}})^T \mathbf{d}$ is zero. So $\nabla f(\tilde{\mathbf{w}})$ is perpendicular to every vector perpendicular to $\nabla h(\tilde{\mathbf{w}})$, i.e. parallel to it: $\nabla f(\tilde{\mathbf{w}}) = -\tilde{\beta} \nabla h(\tilde{\mathbf{w}})$. Sufficiency: with f convex and h affine, $L(\cdot, \tilde{\beta})$ is convex, so $\nabla_{\mathbf{w}} L = 0$ makes $\tilde{\mathbf{w}}$ its global minimum (§2.5); for any feasible $\mathbf{w}$, $f(\mathbf{w}) = L(\mathbf{w}, \tilde{\beta}) \geq L(\tilde{\mathbf{w}}, \tilde{\beta}) = f(\tilde{\mathbf{w}})$. $\square$

The slides call (10) “necessary and sufficient”; without convexity it is only necessary: it finds all *candidates* (minima, maxima, saddle points), and we compare their f values, as in example (c) below. The sign of β is not restricted for equality constraints, and the same conditions apply to a maximization.

4.3 Worked Examples

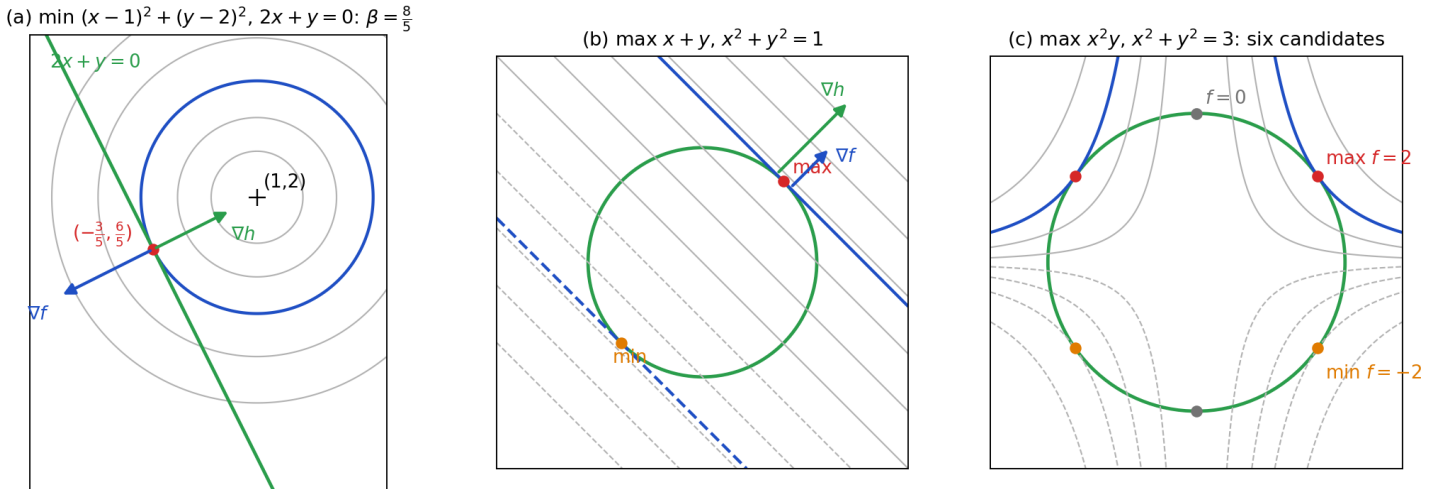


Figure 7: Lagrange multipliers at work: at each solution a level curve of f (blue) is tangent to the constraint (green), so $\nabla f \parallel \nabla h$.

(a) Closest point on a line. Minimize $f(x, y) = (x - 1)^2 + (y - 2)^2$ subject to $h(x, y) = 2x + y = 0$. With $L = (x - 1)^2 + (y - 2)^2 + \beta(2x + y)$:

$$\frac{\partial L}{\partial x} = 2(x - 1) + 2\beta = 0 \Rightarrow x = 1 - \beta; \quad \frac{\partial L}{\partial y} = 2(y - 2) + \beta = 0 \Rightarrow y = 2 - \frac{\beta}{2}; \quad (11)$$

$$\frac{\partial L}{\partial \beta} = 2x + y = 2(1 - \beta) + 2 - \frac{\beta}{2} = 0 \Rightarrow \beta = \frac{8}{5}, \quad x = -\frac{3}{5}, \quad y = \frac{6}{5}, \quad f = \frac{16}{5}. \quad (12)$$

Check: the level curves of f are circles around $(1, 2)$; the vector from the solution to the center, $(1, 2) - (-\frac{3}{5}, \frac{6}{5}) = (\frac{8}{5}, \frac{4}{5})$, is parallel to the line’s normal $(2, 1)$, i.e. perpendicular to the line (Figure 7a).

(b) **Maximize** $f = x + y$ **on the unit circle** $h = x^2 + y^2 - 1 = 0$. $L = x + y + \beta(x^2 + y^2 - 1)$: $1 + 2\beta x = 0$ and $1 + 2\beta y = 0$, so $x = y = -\frac{1}{2\beta}$; the constraint gives $\frac{1}{2\beta^2} = 1$, $\beta = \pm\frac{1}{\sqrt{2}}$. The two candidates are $x = y = \pm\frac{1}{\sqrt{2}}$; the maximum is $f = \sqrt{2}$ at $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ ($\beta = -\frac{1}{\sqrt{2}}$), the minimum $-\sqrt{2}$ at the opposite point (Figure 7b).

(c) **Maximize** $f = x^2y$ **on the circle** $x^2 + y^2 = 3$. $L = x^2y + \beta(x^2 + y^2 - 3)$:

$$\frac{\partial L}{\partial x} = 2x(y + \beta) = 0, \quad \frac{\partial L}{\partial y} = x^2 + 2\beta y = 0. \quad (13)$$

Either $x = 0$, giving $y = \pm\sqrt{3}$ and $f = 0$; or $y = -\beta$, and then $x^2 = 2y^2$, so $3y^2 = 3$: $y = \pm 1$, $x = \pm\sqrt{2}$, $f = 2y$. Six candidates; comparing values, the maximum is $f = 2$ at $(\pm\sqrt{2}, 1)$, the minimum -2 at $(\pm\sqrt{2}, -1)$, and $(0, \pm\sqrt{3})$ are neither (Figure 7c).

4.4 Connection: PCA

The PCA note derived the principal directions exactly this way: maximize $\mathbf{e}^T \Sigma \mathbf{e}$ subject to $\mathbf{e}^T \mathbf{e} = 1$. With $L = \mathbf{e}^T \Sigma \mathbf{e} - \lambda(\mathbf{e}^T \mathbf{e} - 1)$, $\nabla_{\mathbf{e}} L = 2\Sigma \mathbf{e} - 2\lambda \mathbf{e} = 0$ gives $\Sigma \mathbf{e} = \lambda \mathbf{e}$, and the objective at the solution is $\mathbf{e}^T \Sigma \mathbf{e} = \lambda$.

Key point: A Lagrange multiplier is not just a bookkeeping device: it often has a meaning. In PCA it *is* the eigenvalue (the variance captured). In general $\tilde{\beta}_j$ measures how fast the optimal value changes if constraint j is relaxed.

5 Inequality Constraints and the KKT Conditions

Now the general problem:

$$\min_{\mathbf{w}} f(\mathbf{w}) \quad \text{subject to} \quad g_i(\mathbf{w}) \leq 0, \quad i = 1..k, \quad h_j(\mathbf{w}) = 0, \quad j = 1..m. \quad (14)$$

It is a **convex problem** if f and all g_i are convex and all h_j are affine. The Lagrangian gets a multiplier for each constraint:

$$L(\mathbf{w}, \alpha, \beta) = f(\mathbf{w}) + \sum_{i=1}^k \alpha_i g_i(\mathbf{w}) + \sum_{j=1}^m \beta_j h_j(\mathbf{w}), \quad \alpha_i \geq 0. \quad (15)$$

5.1 Active and Inactive Constraints

At the solution each inequality constraint is either

- **inactive**, $g_i(\tilde{\mathbf{w}}) < 0$: it does not touch the solution, removing it changes nothing, and its multiplier is $\alpha_i = 0$ (Figure 8a); or
- **active**, $g_i(\tilde{\mathbf{w}}) = 0$: the solution sits on its boundary and it behaves like an equality constraint, $\nabla f = -\alpha_i \nabla g_i$ (Figure 8b).

Both cases are captured by one equation, **complementary slackness**: $\alpha_i g_i(\tilde{\mathbf{w}}) = 0$ (at least one of the two factors is zero).

Why $\alpha_i \geq 0$. ∇g_i points to where g_i grows, i.e. *out* of the feasible set. If $\nabla f = -\alpha \nabla g$ with $\alpha \geq 0$, the descent direction $-\nabla f$ points outward too: every direction $\mathbf{d}$ that stays feasible ($\nabla g^T \mathbf{d} \leq 0$) has $\nabla f^T \mathbf{d} = -\alpha \nabla g^T \mathbf{d} \geq 0$, so no feasible move decreases f . If α were negative, f would decrease by moving inward and the minimum would not be on the boundary.

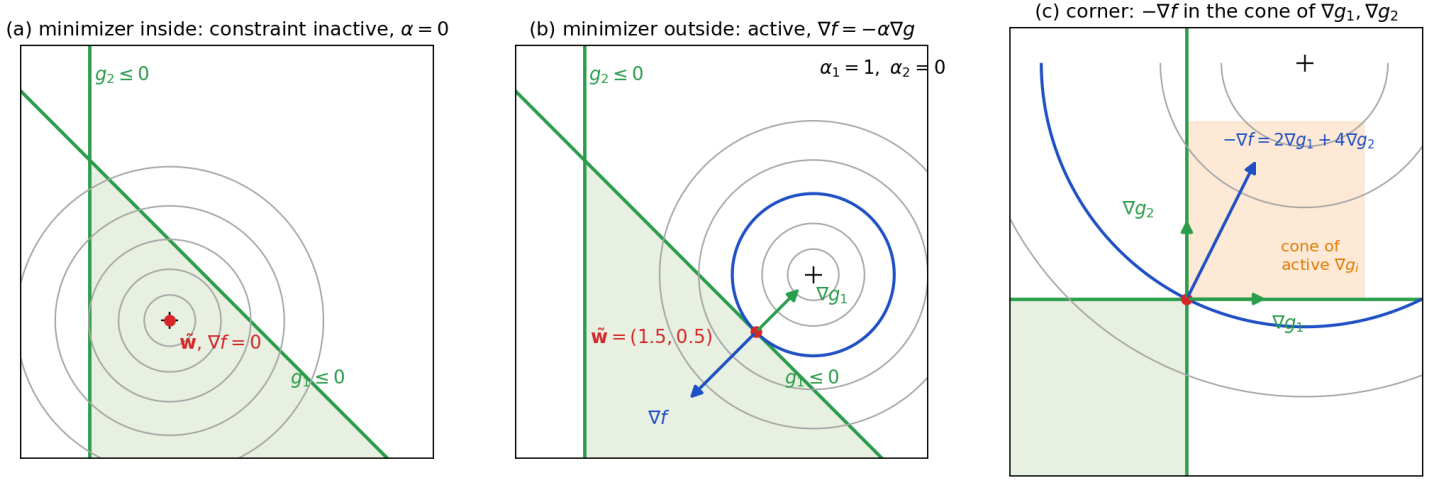


Figure 8: Feasible set (green) $g_1 = w_1 + w_2 - 2 \leq 0$, $g_2 = -w_1 \leq 0$. (a) The unconstrained minimizer is feasible: no constraint matters, all $\alpha_i = 0$. (b) The unconstrained minimizer $(2, 1)$ is infeasible: the solution is on the boundary of g_1 , where $\nabla f = -\alpha_1 \nabla g_1$ with $\alpha_1 = 1 \geq 0$. (c) At a corner two constraints are active and $-\nabla f$ lies in the cone spanned by their gradients.

5.2 Several Active Constraints

At a corner (Figure 8c: minimize $(w_1 - 2)^2 + (w_2 - 3)^2$ subject to $w_1 \leq 1$, $w_2 \leq 1$, solution $(1, 1)$) both constraints are active and

$$-\nabla f(\tilde{\mathbf{w}}) = \sum_{i \text{ active}} \alpha_i \nabla g_i(\tilde{\mathbf{w}}), \quad \alpha_i \geq 0: \quad (2, 4)^T = 2(1, 0)^T + 4(0, 1)^T. \quad (16)$$

Geometrically, $-\nabla f$ must lie in the **cone** spanned by the active constraint gradients; otherwise there would be a feasible direction that decreases f .

5.3 The KKT Conditions

Karush–Kuhn–Tucker (KKT) conditions for problem (14)

stationarity:	$\nabla_{\mathbf{w}} L(\tilde{\mathbf{w}}, \tilde{\alpha}, \tilde{\beta}) = \nabla f(\tilde{\mathbf{w}}) + \sum_i \tilde{\alpha}_i \nabla g_i(\tilde{\mathbf{w}}) + \sum_j \tilde{\beta}_j \nabla h_j(\tilde{\mathbf{w}}) = 0$
primal feasibility:	$g_i(\tilde{\mathbf{w}}) \leq 0$ for all i , $h_j(\tilde{\mathbf{w}}) = 0$ for all j
dual feasibility:	$\tilde{\alpha}_i \geq 0$ for all i
complementary slackness:	$\tilde{\alpha}_i g_i(\tilde{\mathbf{w}}) = 0$ for all i

For a convex problem the KKT conditions are **sufficient** for optimality (Theorem in §7.2), and, under a mild “constraint qualification” such as Slater’s condition, also **necessary** (§7.3–§7.5). The old notes listed $\nabla_{\alpha} L = 0$ among the conditions; that would force $g_i(\tilde{\mathbf{w}}) = 0$ for every i , i.e. all constraints active. The correct conditions are primal feasibility plus complementary slackness.

Worked example (Figure 8b). Minimize $f = (w_1 - 2)^2 + (w_2 - 1)^2$ subject to $g_1 = w_1 + w_2 - 2 \leq 0$ and $g_2 = -w_1 \leq 0$. Solve by guessing which constraints are active:

- *Neither active* ($\alpha_1 = \alpha_2 = 0$): $\nabla f = 0$ gives $(2, 1)$, but $g_1 = 1 > 0$: infeasible, reject.
- *Only g_1 active* ($\alpha_2 = 0$, $w_1 + w_2 = 2$): stationarity $2(w_1 - 2) + \alpha_1 = 0$, $2(w_2 - 1) + \alpha_1 = 0$ gives $w_1 - 2 = w_2 - 1$, so $w_1 = 1.5$, $w_2 = 0.5$, $\alpha_1 = 2(2 - 1.5) = 1$.

- Check: $\alpha_1 = 1 \geq 0$ ✓; $g_2 = -1.5 < 0$ ✓ (inactive, $\alpha_2 = 0$) ✓. All KKT conditions hold, and the problem is convex, so $\tilde{\mathbf{w}} = (1.5, 0.5)$ with $p^* = 0.5$ is the global solution.

Remark. The slides turn each equality $h_j = 0$ into two inequalities $h_j \leq 0$ and $-h_j \leq 0$, so only inequalities remain. Their two multipliers $\alpha^+, \alpha^- \geq 0$ enter the Lagrangian as $(\alpha^+ - \alpha^-)h_j$; the difference $\beta_j = \alpha^+ - \alpha^-$ can have any sign. That is why we keep equality multipliers β_j free of sign instead.

6 Lagrangian Duality

6.1 The Primal as a Min–Max

For a fixed $\mathbf{w}$, maximize the Lagrangian (15) over the multipliers:

$$\max_{\alpha \geq 0, \beta} L(\mathbf{w}, \alpha, \beta) = \begin{cases} f(\mathbf{w}) & \text{if } \mathbf{w} \text{ is feasible} \\ +\infty & \text{if some } g_i(\mathbf{w}) > 0 \text{ or } h_j(\mathbf{w}) \neq 0. \end{cases} \quad (17)$$

(If $g_i(\mathbf{w}) > 0$, send $\alpha_i \rightarrow \infty$; if $h_j(\mathbf{w}) \neq 0$, send $\beta_j \rightarrow \pm\infty$; if feasible, the best is $\alpha_i g_i = 0$.) So the constrained problem is an unconstrained **min–max**: $p^* = \min_{\mathbf{w}} \max_{\alpha \geq 0, \beta} L$. Swapping the order gives the **dual problem**.

6.2 The Dual Function and Weak Duality

$$\theta(\alpha, \beta) = \min_{\mathbf{w}} L(\mathbf{w}, \alpha, \beta), \quad d^* = \max_{\alpha \geq 0, \beta} \theta(\alpha, \beta). \quad (18)$$

(The SVM Advanced note writes θ_D .) Two facts hold for *every* problem, convex or not:

- **θ is concave.** For each fixed $\mathbf{w}$, $L(\mathbf{w}, \alpha, \beta)$ is affine in (α, β) ; θ is the pointwise minimum of these affine functions, and a minimum of affine functions is concave. So the dual is always a concave maximization (a convex problem), even if the primal is not.
- **Weak duality:** $d^* \leq p^*$. For any feasible $\mathbf{w}$ and any $\alpha \geq 0, \beta$:

$$\theta(\alpha, \beta) \leq L(\mathbf{w}, \alpha, \beta) = f(\mathbf{w}) + \underbrace{\sum_i \alpha_i g_i(\mathbf{w})}_{\leq 0} + \underbrace{\sum_j \beta_j h_j(\mathbf{w})}_{=0} \leq f(\mathbf{w}). \quad (19)$$

Every dual value is a lower bound on every feasible primal value.

The difference $p^* - d^* \geq 0$ is the **duality gap**. In the slides' form (“KKT gap”): if $\mathbf{w}$ is feasible and $(\mathbf{w}', \alpha)$ satisfies $\nabla_{\mathbf{w}} L(\mathbf{w}', \alpha) = 0$ with $\alpha \geq 0$ (and L convex in $\mathbf{w}$, so $\mathbf{w}'$ minimizes it), then $f(\mathbf{w}) \geq f(\tilde{\mathbf{w}}) \geq f(\mathbf{w}') + \sum_i \alpha_i g_i(\mathbf{w}')$. The two ends bracket the optimum; when they meet, both are optimal.

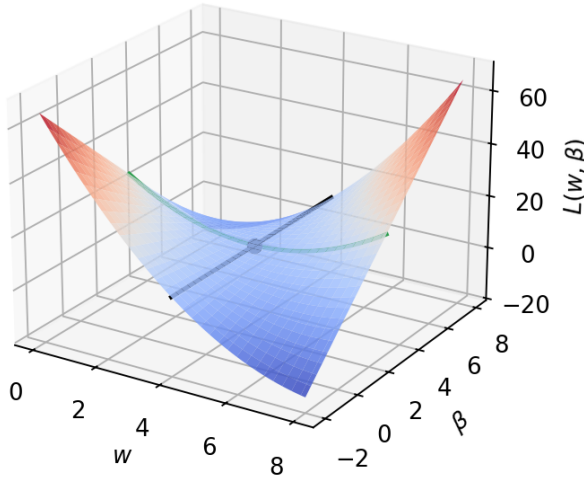
6.3 A Saddle Point, Explicitly

Minimize $f(w) = (w - 7)^2$ subject to $h(w) = 2w - 8 = 0$. The only feasible point is $w = 4$, so $p^* = 9$. The Lagrangian is $L(w, \beta) = (w - 7)^2 + \beta(2w - 8)$. Minimizing over w : $2(w - 7) + 2\beta = 0$, $w = 7 - \beta$, and

$$\theta(\beta) = \beta^2 + \beta(14 - 2\beta - 8) = -\beta^2 + 6\beta. \quad (20)$$

- *Weak duality:* every $\theta(\beta) \leq 9$; e.g. $\theta(0) = 0$ is the unconstrained minimum, a poor lower bound.

(a) $L(w, \beta) = (w - 7)^2 + \beta(2w - 8)$: saddle at $(4, 3)$



(b) every $\theta(\beta) \leq p^*$; the best one touches it

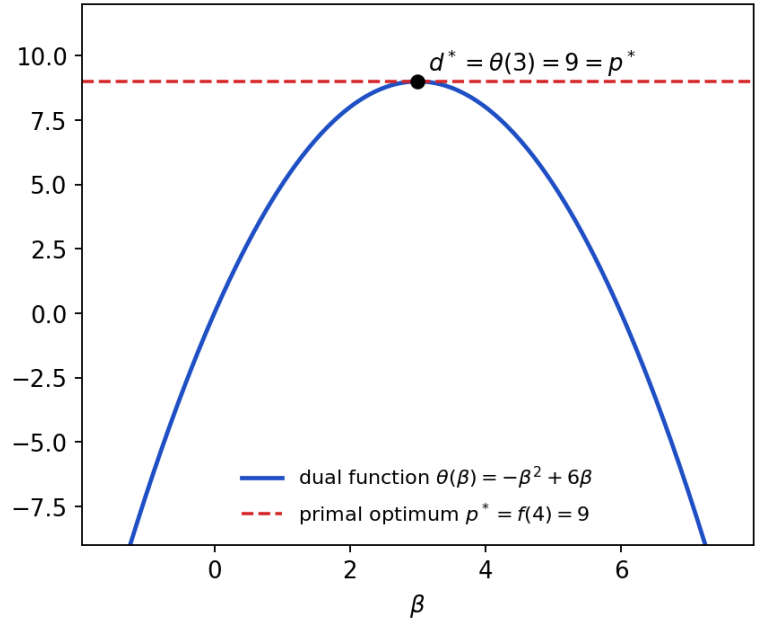


Figure 9: (a) The Lagrangian of: minimize $(w - 7)^2$ subject to $2w - 8 = 0$. Along $\beta = 3$ (green) it is a parabola in w with minimum at $w = 4$; along $w = 4$ (black) it is constant, because $h(4) = 0$. (b) The dual function: always below p^* , touching it at $\beta = 3$.

- *Strong duality*: the maximum is at $\beta = 3$, $d^* = \theta(3) = 9 = p^*$, with $w = 7 - 3 = 4$.
- *Saddle point*: $(\tilde{w}, \tilde{\beta}) = (4, 3)$ is a minimum of L in w ($L(w, 3) = (w - 4)^2 + 9$) and a maximum in β ($L(4, \beta) = 9$), Figure 9.

6.4 The Geometry of Duality

Take one inequality constraint and map every $\mathbf{w}$ to the point $(u, t) = (g(\mathbf{w}), f(\mathbf{w}))$ (Figure 10). The primal optimum p^* is the lowest t among points with $u \leq 0$. For $\alpha \geq 0$,

$$\theta(\alpha) = \min_{\mathbf{w}} (f(\mathbf{w}) + \alpha g(\mathbf{w})) = \min_{(u, t)} (t + \alpha u) : \quad (21)$$

slide the line $t + \alpha u = c$ (slope $-\alpha$) up until it touches the set from below; $\theta(\alpha)$ is where that line crosses $u = 0$. Since points with $u \leq 0$ lie above the line, its height at $u = 0$ is at most p^* : weak duality, as a picture. The dual problem tilts the line to make that height as large as possible. When the set (more precisely its “upper-right hull” $\mathcal{A}$, shaded) is convex, some line touches it exactly at $(0, p^*)$: **no gap**. When it is not, the best line can get stuck below: a positive gap.

Numbers. (a) $f = w^2$, $g = 1 - w$: $\theta(\alpha) = \min_w w^2 + \alpha(1 - w) = \alpha - \frac{\alpha^2}{4}$, maximized at $\alpha = 2$ with $d^* = 1 = p^*$ ($w = 1$). (b) $f = (w^2 - 1)^2 - 0.3w$, $g = w \leq 0$: the feasible minimum is $p^* = 0.294$ at $w = -0.96$, but $d^* = 0$ (at $\alpha = 0.3$, the line through both dips).

6.5 Strong Duality and Slater’s Condition

Theorem (strong duality). *If problem (14) is convex and **Slater’s condition** holds — there is a strictly feasible point, $g_i(\mathbf{w}) < 0$ for all i and $h_j(\mathbf{w}) = 0$ for all j — then $d^* = p^*$, and (if p^* is finite) the dual optimum is attained.*

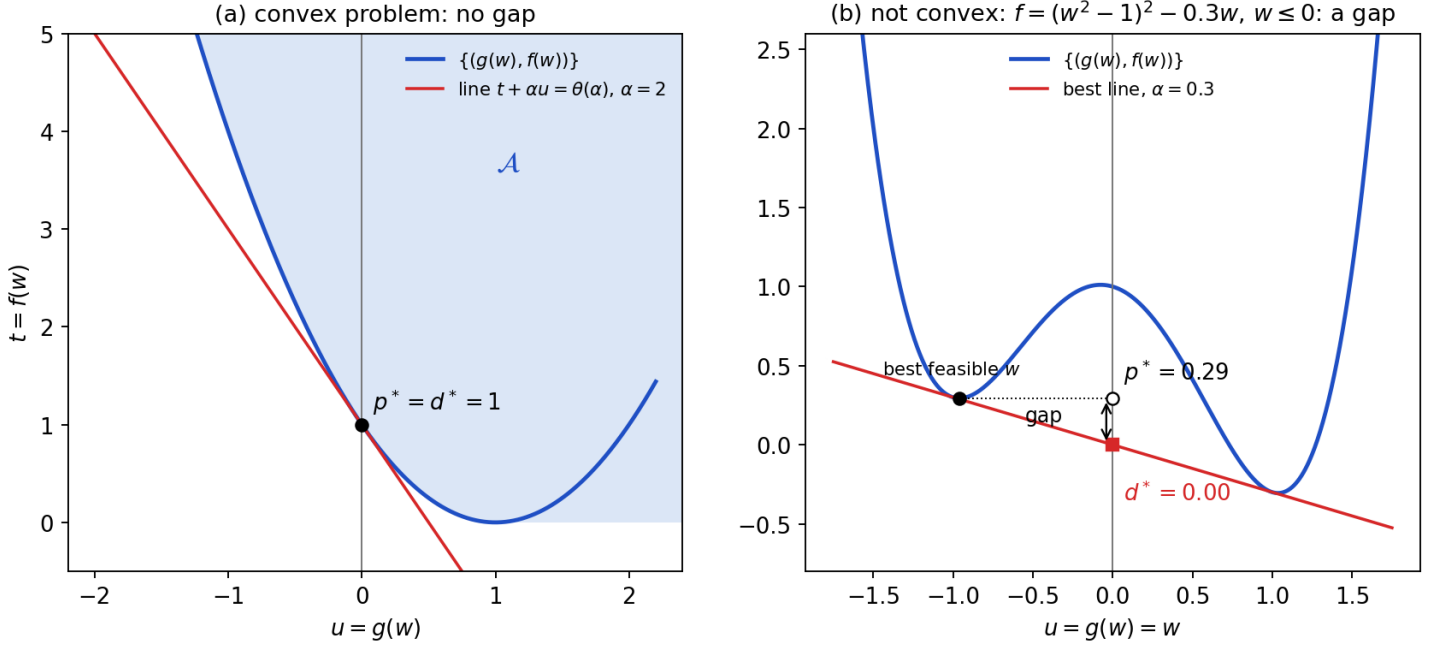


Figure 10: Each $\mathbf{w}$ is plotted at $(u, t) = (g(\mathbf{w}), f(\mathbf{w}))$; feasible points have $u \leq 0$. A dual value $\theta(\alpha)$ is the height at $u = 0$ of a line of slope $-\alpha$ that stays below all the points. (a) Minimize w^2 s.t. $1 - w \leq 0$: the set is convex and the best line (slope -2) passes through $(0, p^*)$, no gap. (b) A non-convex problem: the best line ($\alpha = 0.3$) rests on both dips and crosses $u = 0$ below p^* : $d^* = 0 < p^* = 0.29$.

Proof in §7.3. In practice we then solve whichever of the two problems is easier, and read the other solution off the KKT conditions.

6.6 Worked Dual: Projection onto a Half-Space

Minimize $\frac{1}{2}\|\mathbf{w}\|^2$ subject to $\mathbf{a}\mathbf{w} \geq b$ (a row vector $\mathbf{a}$, $b > 0$): the point of the half-space closest to the origin. Write the constraint as $g(\mathbf{w}) = b - \mathbf{a}\mathbf{w} \leq 0$:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + \alpha(b - \mathbf{a}\mathbf{w}), \quad \nabla_{\mathbf{w}}L = \mathbf{w} - \alpha\mathbf{a}^T = 0 \Rightarrow \mathbf{w} = \alpha\mathbf{a}^T, \quad (22)$$

$$\theta(\alpha) = \frac{1}{2}\alpha^2\|\mathbf{a}\|^2 + \alpha b - \alpha^2\|\mathbf{a}\|^2 = \alpha b - \frac{1}{2}\alpha^2\|\mathbf{a}\|^2, \quad \tilde{\alpha} = \frac{b}{\|\mathbf{a}\|^2}, \quad d^* = \frac{b^2}{2\|\mathbf{a}\|^2}. \quad (23)$$

So $\tilde{\mathbf{w}} = \frac{b}{\|\mathbf{a}\|^2}\mathbf{a}^T$, with $\frac{1}{2}\|\tilde{\mathbf{w}}\|^2 = \frac{b^2}{2\|\mathbf{a}\|^2} = d^*$ (no gap; Slater holds), and the distance from the origin to the half-space is $\|\tilde{\mathbf{w}}\| = b/\|\mathbf{a}\|$. With $\mathbf{a} = (3, 4)$, $b = 10$: $\tilde{\alpha} = 0.4$, $\tilde{\mathbf{w}} = (1.2, 1.6)^T$, distance 2. Notice the pattern: the primal solution is the multiplier times the constraint's “data” $\mathbf{a}^T$. The SVM note applies exactly this to one such constraint per training point, $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i^T$.

7 Theorems and Proofs

More mathematical; skip on first reading. Throughout, problem (14) with Lagrangian (15). To keep the notation light, the proofs treat inequality constraints only; affine equality constraints work the same way (they are two inequalities each, or add a β -coordinate), and we write $\mathbf{g}(\mathbf{w}) = (g_1(\mathbf{w}), \dots, g_k(\mathbf{w}))^T$.

7.1 Saddle Points Are Solutions

$(\tilde{\mathbf{w}}, \tilde{\alpha})$ with $\tilde{\alpha} \geq 0$ is a **saddle point** of L if for all $\mathbf{w}$ and all $\alpha \geq 0$

$$L(\tilde{\mathbf{w}}, \alpha) \leq L(\tilde{\mathbf{w}}, \tilde{\alpha}) \leq L(\mathbf{w}, \tilde{\alpha}). \quad (24)$$

Theorem (saddle point $\Rightarrow$ optimum). *If $(\tilde{\mathbf{w}}, \tilde{\alpha})$ is a saddle point, then $\tilde{\mathbf{w}}$ solves the problem and $\tilde{\alpha}_i g_i(\tilde{\mathbf{w}}) = 0$ for all i . No convexity is needed.*

Proof. The left inequality says $\sum_i \alpha_i g_i(\tilde{\mathbf{w}}) \leq \sum_i \tilde{\alpha}_i g_i(\tilde{\mathbf{w}})$ for all $\alpha \geq 0$. If some $g_i(\tilde{\mathbf{w}}) > 0$, letting $\alpha_i \rightarrow \infty$ violates it; so $\tilde{\mathbf{w}}$ is feasible. With $\alpha = 0$ it gives $\sum_i \tilde{\alpha}_i g_i(\tilde{\mathbf{w}}) \geq 0$; each term is ≤ 0 (as $\tilde{\alpha}_i \geq 0$, $g_i(\tilde{\mathbf{w}}) \leq 0$), so each is 0: complementary slackness. Then for any feasible $\mathbf{w}$, by the right inequality, $f(\tilde{\mathbf{w}}) = L(\tilde{\mathbf{w}}, \tilde{\alpha}) \leq L(\mathbf{w}, \tilde{\alpha}) = f(\mathbf{w}) + \sum_i \tilde{\alpha}_i g_i(\mathbf{w}) \leq f(\mathbf{w})$. $\square$

7.2 KKT Is Sufficient for Convex Problems

Theorem (KKT sufficiency). *If f and all g_i are convex and differentiable and $(\tilde{\mathbf{w}}, \tilde{\alpha})$ satisfies the KKT conditions, then $\tilde{\mathbf{w}}$ is a global solution.*

Proof. For any feasible $\mathbf{w}$, using (4) for f , then stationarity, then (4) for each g_i (multiplied by $-\tilde{\alpha}_i \leq 0$, which flips the inequality), then complementary slackness:

$$\begin{aligned} f(\mathbf{w}) - f(\tilde{\mathbf{w}}) &\geq \nabla f(\tilde{\mathbf{w}})^T (\mathbf{w} - \tilde{\mathbf{w}}) = - \sum_i \tilde{\alpha}_i \nabla g_i(\tilde{\mathbf{w}})^T (\mathbf{w} - \tilde{\mathbf{w}}) \\ &\geq - \sum_i \tilde{\alpha}_i (g_i(\mathbf{w}) - g_i(\tilde{\mathbf{w}})) = - \sum_i \tilde{\alpha}_i g_i(\mathbf{w}) \geq 0. \end{aligned} \quad (25)$$

The last step uses $\tilde{\alpha}_i \geq 0$ and $g_i(\mathbf{w}) \leq 0$. $\square$

7.3 Strong Duality under Slater's Condition

Theorem (strong duality). *If f and all g_i are convex, p^* is finite, and some $\bar{\mathbf{w}}$ has $g_i(\bar{\mathbf{w}}) < 0$ for all i , then there is $\tilde{\alpha} \geq 0$ with $\theta(\tilde{\alpha}) = p^*$; hence $d^* = p^*$ and the dual optimum is attained.*

Proof. Define the sets (Figure 10a shows $\mathcal{A}$ for one constraint)

$$\mathcal{A} = \{(\mathbf{u}, t) : \exists \mathbf{w} \text{ with } \mathbf{g}(\mathbf{w}) \leq \mathbf{u}, f(\mathbf{w}) \leq t\}, \quad \mathcal{B} = \{(\mathbf{0}, s) : s < p^*\}. \quad (26)$$

$\mathcal{A}$ is convex: if $(\mathbf{u}_1, t_1)$ and $(\mathbf{u}_2, t_2)$ are witnessed by $\mathbf{w}_1, \mathbf{w}_2$, then $\mathbf{w}_\lambda = \lambda \mathbf{w}_1 + (1 - \lambda) \mathbf{w}_2$ witnesses their convex combination, by convexity of f and the g_i . $\mathcal{B}$ is convex, and $\mathcal{A} \cap \mathcal{B} = \emptyset$: a point $(\mathbf{0}, s) \in \mathcal{A}$ would be a feasible $\mathbf{w}$ with $f(\mathbf{w}) \leq s < p^*$. By the *separating hyperplane theorem* there are $(\tilde{\alpha}, \mu) \neq 0$ and c with

$$\tilde{\alpha}^T \mathbf{u} + \mu t \geq c \text{ on } \mathcal{A}, \quad \tilde{\alpha}^T \mathbf{u} + \mu t \leq c \text{ on } \mathcal{B}. \quad (27)$$

Since $\mathcal{A}$ is unbounded upward in every u_i and in t , the left inequality forces $\tilde{\alpha} \geq 0$ and $\mu \geq 0$. On $\mathcal{B}$ it says $\mu s \leq c$ for all $s < p^*$, so $\mu p^* \leq c$. Taking $(\mathbf{g}(\mathbf{w}), f(\mathbf{w})) \in \mathcal{A}$:

$$\sum_i \tilde{\alpha}_i g_i(\mathbf{w}) + \mu f(\mathbf{w}) \geq \mu p^* \quad \text{for all } \mathbf{w}. \quad (28)$$

Case $\mu > 0$. Divide by μ : $L(\mathbf{w}, \tilde{\alpha}/\mu) \geq p^*$ for all $\mathbf{w}$, so $\theta(\tilde{\alpha}/\mu) \geq p^*$; with weak duality, equality. Case $\mu = 0$. Then $\tilde{\alpha} \neq 0$, $\tilde{\alpha} \geq 0$, and (28) gives $\sum_i \tilde{\alpha}_i g_i(\mathbf{w}) \geq 0$ for all $\mathbf{w}$. But at the Slater point every $g_i(\bar{\mathbf{w}}) < 0$, so $\sum_i \tilde{\alpha}_i g_i(\bar{\mathbf{w}}) < 0$: a contradiction. So $\mu > 0$. $\square$

Geometrically, the separating hyperplane is the supporting line of Figure 10a through $(0, p^*)$; Slater's condition makes sure it is not vertical ($\mu \neq 0$), i.e. that its slope $-\alpha$ is finite.

7.4 KKT Is Necessary; Saddle Points Exist

Theorem (KKT necessity). *Under the assumptions of the strong duality theorem, with f, g_i differentiable: if $\tilde{\mathbf{w}}$ is a solution, there is $\tilde{\alpha} \geq 0$ such that $(\tilde{\mathbf{w}}, \tilde{\alpha})$ satisfies the KKT conditions and is a saddle point of L .*

Proof. Take $\tilde{\alpha}$ from the strong duality theorem, $\theta(\tilde{\alpha}) = p^* = f(\tilde{\mathbf{w}})$. Then

$$f(\tilde{\mathbf{w}}) = \theta(\tilde{\alpha}) = \min_{\mathbf{w}} L(\mathbf{w}, \tilde{\alpha}) \leq L(\tilde{\mathbf{w}}, \tilde{\alpha}) = f(\tilde{\mathbf{w}}) + \sum_i \tilde{\alpha}_i g_i(\tilde{\mathbf{w}}) \leq f(\tilde{\mathbf{w}}). \quad (29)$$

So both inequalities are equalities. The second gives $\sum_i \tilde{\alpha}_i g_i(\tilde{\mathbf{w}}) = 0$, and since every term is ≤ 0 , complementary slackness. The first says $\tilde{\mathbf{w}}$ minimizes $L(\cdot, \tilde{\alpha})$, so $\nabla_{\mathbf{w}} L(\tilde{\mathbf{w}}, \tilde{\alpha}) = 0$: stationarity. Feasibility of $\tilde{\mathbf{w}}$ and $\tilde{\alpha} \geq 0$ hold by construction. For the saddle point: $L(\tilde{\mathbf{w}}, \tilde{\alpha}) = \min_{\mathbf{w}} L(\mathbf{w}, \tilde{\alpha}) \leq L(\mathbf{w}, \tilde{\alpha})$, and for any $\alpha \geq 0$, $L(\tilde{\mathbf{w}}, \alpha) = f(\tilde{\mathbf{w}}) + \sum_i \alpha_i g_i(\tilde{\mathbf{w}}) \leq f(\tilde{\mathbf{w}}) = L(\tilde{\mathbf{w}}, \tilde{\alpha})$. $\square$

7.5 Constraint Qualifications

Necessity needs some condition on the constraints, a **constraint qualification**. Without one, KKT can fail even at the true optimum:

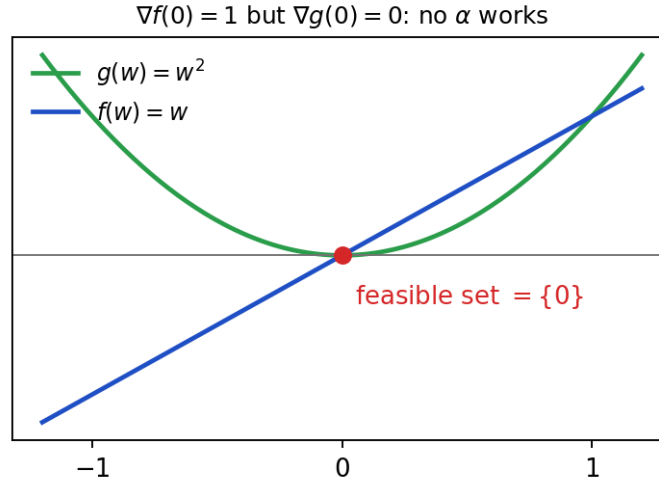


Figure 11: Minimize $f(w) = w$ subject to $g(w) = w^2 \leq 0$. The only feasible point is $w = 0$, but stationarity $1 + 2\alpha w = 0$ has no solution there.

Counterexample. Minimize $f(w) = w$ subject to $g(w) = w^2 \leq 0$. The feasible set is $\{0\}$, so $\tilde{w} = 0$ and $p^* = 0$. Stationarity requires $f'(0) + \alpha g'(0) = 1 + 0 = 0$: impossible for every α (Figure 11). The dual function is $\theta(\alpha) = \min_w (w + \alpha w^2) = -\frac{1}{4\alpha}$, which approaches $0 = p^*$ as $\alpha \rightarrow \infty$ but never reaches it: no dual optimum, no KKT pair. Slater fails because no w has $w^2 < 0$.

Common qualifications.

- **Slater** (convex problems): some $\mathbf{w}$ has $g_i(\mathbf{w}) < 0$ for all i (*strict*; the slides wrote $\leq$, which every feasible point satisfies and which the counterexample above also satisfies) and $h_j(\mathbf{w}) = 0$ with h_j affine.
- **Karlin's condition** (equivalent to Slater for convex g_i): for every $\alpha \geq 0$, $\alpha \neq 0$, some $\mathbf{w}$ has $\sum_i \alpha_i g_i(\mathbf{w}) < 0$. (The slides' version, with $\alpha \in [0, 1]^k$ and " $= 0$ ", is a garbled form of this.) Slater $\Rightarrow$ Karlin is immediate: at the Slater point every term is ≤ 0 and at least one, with $\alpha_i > 0$, is < 0 . The converse follows from a separating-hyperplane argument like the one above.
- **Strict convexity** (from the slides): if the feasible set contains two distinct points $\mathbf{w}_1 \neq \mathbf{w}_2$ and all g_i are strictly convex, Slater holds: the midpoint has $g_i\left(\frac{\mathbf{w}_1 + \mathbf{w}_2}{2}\right) < \frac{1}{2}g_i(\mathbf{w}_1) + \frac{1}{2}g_i(\mathbf{w}_2) \leq 0$.
- **LICQ** (any problem, convex or not): the gradients of the active inequality constraints and of the equality constraints are linearly independent at $\tilde{\mathbf{w}}$. Then every local minimum satisfies KKT (this generalizes the regular-point condition of Lagrange's theorem).

7.6 Summary of the Relations

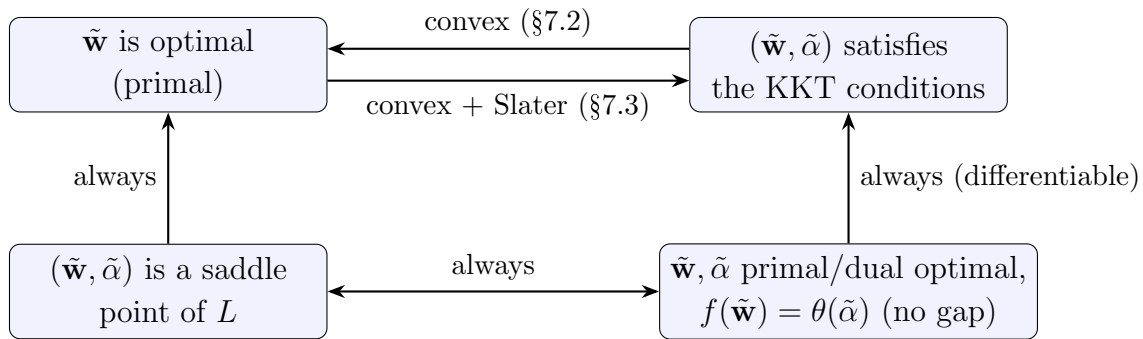


Figure 12: How the four notions relate. For a convex, differentiable problem satisfying Slater's condition all four are equivalent.

The saddle-point and zero-gap statements are the same thing written two ways: $(\tilde{\mathbf{w}}, \tilde{\alpha})$ is a saddle point exactly when $\tilde{\mathbf{w}}$ is primal optimal, $\tilde{\alpha}$ is dual optimal, and $p^* = d^*$ (the arguments above show both directions). With zero gap the KKT conditions follow (proof of KKT necessity), and KKT gives optimality for convex problems.

8 Algorithms for Constrained Problems

8.1 Projected Gradient Descent

When projecting onto the feasible set C is easy (a ball, a box, a half-space), take a gradient step and then move to the nearest feasible point:

$$\mathbf{w} \leftarrow \Pi_C(\mathbf{w} - \eta \nabla f(\mathbf{w})), \quad \Pi_C(\mathbf{v}) = \arg \min_{\mathbf{w} \in C} \|\mathbf{w} - \mathbf{v}\|. \quad (30)$$

For the ball $\|\mathbf{w}\| \leq r$, $\Pi_C(\mathbf{v}) = \mathbf{v} \cdot \min(1, r/\|\mathbf{v}\|)$; for a box, clip each coordinate (exactly what SMO's "clip α_j to $[L, H]$ " does).

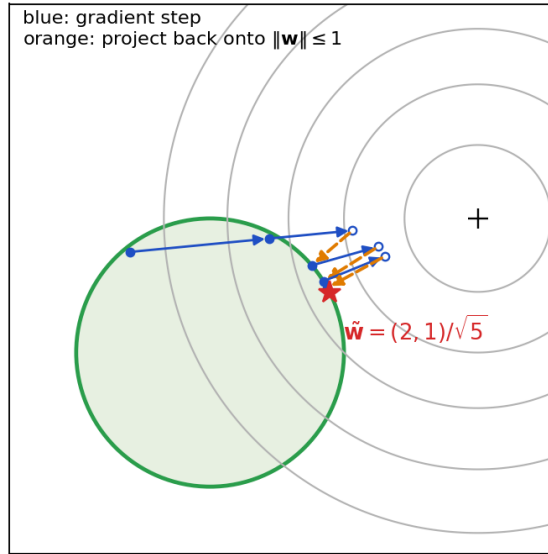


Figure 13: Projected gradient descent for $\min(w_1 - 2)^2 + (w_2 - 1)^2$ s.t. $\|\mathbf{w}\| \leq 1$: a gradient step (blue), then back to the nearest feasible point (orange). It converges to $\frac{1}{\sqrt{5}}(2, 1) = (0.894, 0.447)$.

8.2 Penalty and Barrier Methods; Interior Point

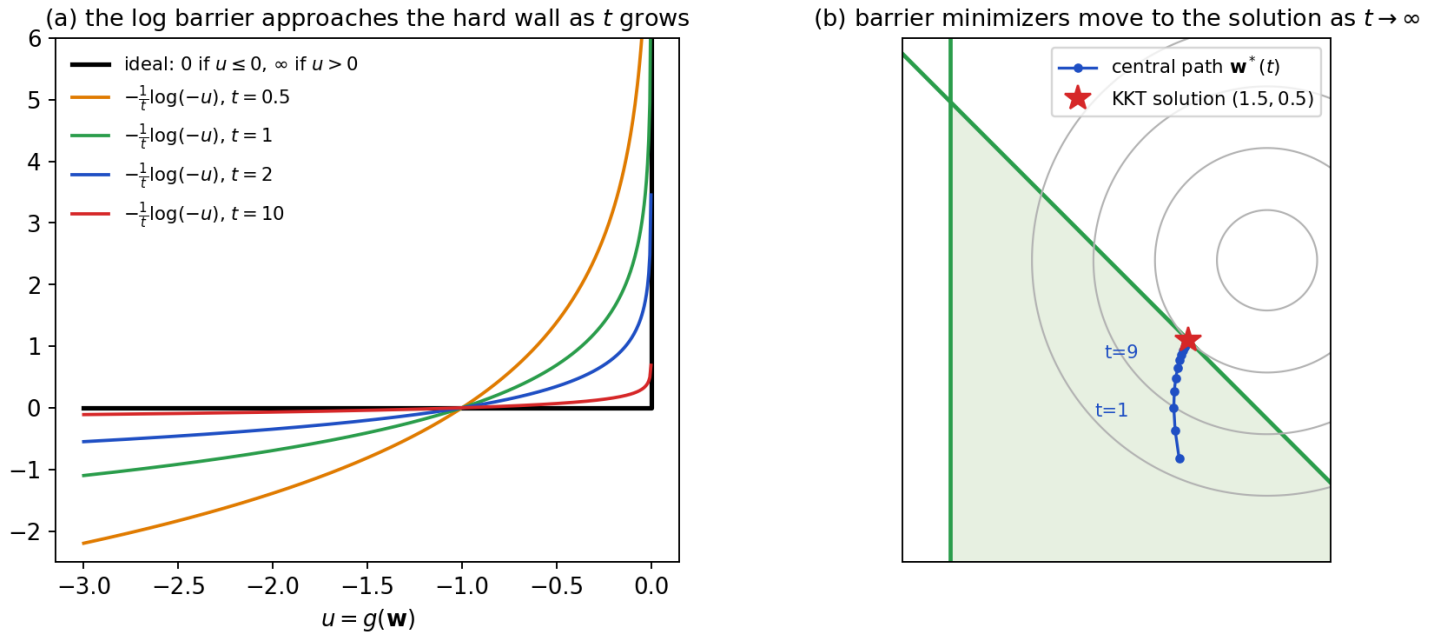


Figure 14: (a) The log barrier $-\frac{1}{t} \log(-u)$ approximates the ideal “wall” (0 inside, ∞ outside) better as t grows. (b) For the example of §5 the barrier minimizers $\mathbf{w}^*(t)$ stay strictly inside and trace the central path to the KKT solution $(1.5, 0.5)$.

Replace the constraints by a term in the objective, then solve unconstrained problems (with Newton’s method):

- **Penalty:** $\min f(\mathbf{w}) + \frac{\mu}{2} \sum_i \max(0, g_i(\mathbf{w}))^2$. Violations are allowed but penalized; the solutions approach the constrained one from *outside* as $\mu \rightarrow \infty$.

- **Log barrier:** $\min \mathbf{f}(\mathbf{w}) - \frac{1}{t} \sum_i \log(-g_i(\mathbf{w}))$. The barrier is finite only strictly inside the feasible set and blows up at its boundary (Figure 14a), so the minimizers $\mathbf{w}^*(t)$ stay *inside*; as $t \rightarrow \infty$ they follow the **central path** to the solution (Figure 14b).

Why the barrier works, via duality. Setting the gradient of the barrier objective to zero gives $\nabla f + \sum_i \alpha_i(t) \nabla g_i = 0$ with $\alpha_i(t) = -\frac{1}{t g_i(\mathbf{w}^*(t))} > 0$. So $\mathbf{w}^*(t)$ minimizes $L(\cdot, \alpha(t))$, and $\theta(\alpha(t)) = f(\mathbf{w}^*(t)) + \sum_i \alpha_i(t) g_i(\mathbf{w}^*(t)) = f(\mathbf{w}^*(t)) - \frac{k}{t}$. By weak duality,

$$f(\mathbf{w}^*(t)) - p^* \leq \frac{k}{t} \quad (k = \text{number of inequality constraints}), \quad (31)$$

a certificate of how close we are, and $\alpha(t)$ converges to the KKT multipliers. In the example ($k = 2$): $t = 10$ gives $\mathbf{w}^* = (1.472, 0.439)$, $\alpha = (1.12, 0.068)$, $f - p^* = 0.094 \leq 0.2$; $t = 100$ gives $(1.4967, 0.4934)$, $\alpha = (1.013, 0.007)$, $f - p^* = 0.0099 \leq 0.02$ — approaching $\alpha = (1, 0)$. **Interior-point methods** do exactly this: Newton's method on the barrier problem, then increase t (e.g. $\times 10$) and repeat from the previous solution.

8.3 Dual Ascent

Since θ is concave, maximize it by (projected) gradient ascent. Its gradient at α is the constraint vector $\mathbf{g}(\mathbf{w}(\alpha))$ at the minimizer $\mathbf{w}(\alpha)$ of $L(\cdot, \alpha)$:

$$\mathbf{w}(\alpha) = \arg \min_{\mathbf{w}} L(\mathbf{w}, \alpha), \quad \alpha \leftarrow \max(0, \alpha + \eta \mathbf{g}(\mathbf{w}(\alpha))). \quad (32)$$

Violated constraints ($g_i > 0$) get a larger price α_i ; satisfied ones relax toward 0. On the example of §5 ($\eta = 0.5$, from $\alpha = 0$) it converges to $\alpha = (1, 0)$, $\mathbf{w} = (1.5, 0.5)$. Solving a dual instead of the primal, often one coordinate pair at a time, is the strategy of SMO.

9 Constraints versus Regularization

Compare a constrained and a regularized least squares problem:

$$(C) \quad \min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 \quad \text{s.t.} \quad \|\mathbf{w}\|^2 \leq r^2 \quad (R) \quad \min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2. \quad (33)$$

The KKT stationarity condition of (C) is $2X^T(X\mathbf{w} - \mathbf{y}) + 2\alpha\mathbf{w} = 0$, i.e. $\mathbf{w} = (X^T X + \alpha I)^{-1} X^T \mathbf{y}$: the **ridge** solution with $\lambda = \alpha$. So the regularization strength *is* the Lagrange multiplier of a norm constraint, and each λ corresponds to the radius $r = \|\mathbf{w}_{\text{ridge}}(\lambda)\|$. (Checked on random data, $N = 30$, $D = 3$, $r = 1$: the constrained solution has multiplier $\alpha = 43.66$, and ridge with $\lambda = 43.66$ returns the same $\mathbf{w}$ to 5 digits.) The same holds for the **lasso** (HW2): $\min \text{loss} + \lambda \|\mathbf{w}\|_1$ corresponds to the constraint $\|\mathbf{w}\|_1 \leq r$. Figure 15 shows why lasso gives **sparse** weights: the L1 ball has corners on the axes, and an elliptical loss contour typically meets it first at a corner, where some coordinates are exactly zero ($\tilde{\mathbf{w}} = (1, 0)$ in the figure), while the round L2 ball has no corners ($\tilde{\mathbf{w}} = (0.89, 0.45)$).

10 Summary

- **Convexity** makes a problem easy: every local minimum is global, and a zero gradient (unconstrained) or the KKT conditions (constrained) certify the global optimum. Check convexity with the rules: sums, affine compositions, maxima of convex functions.

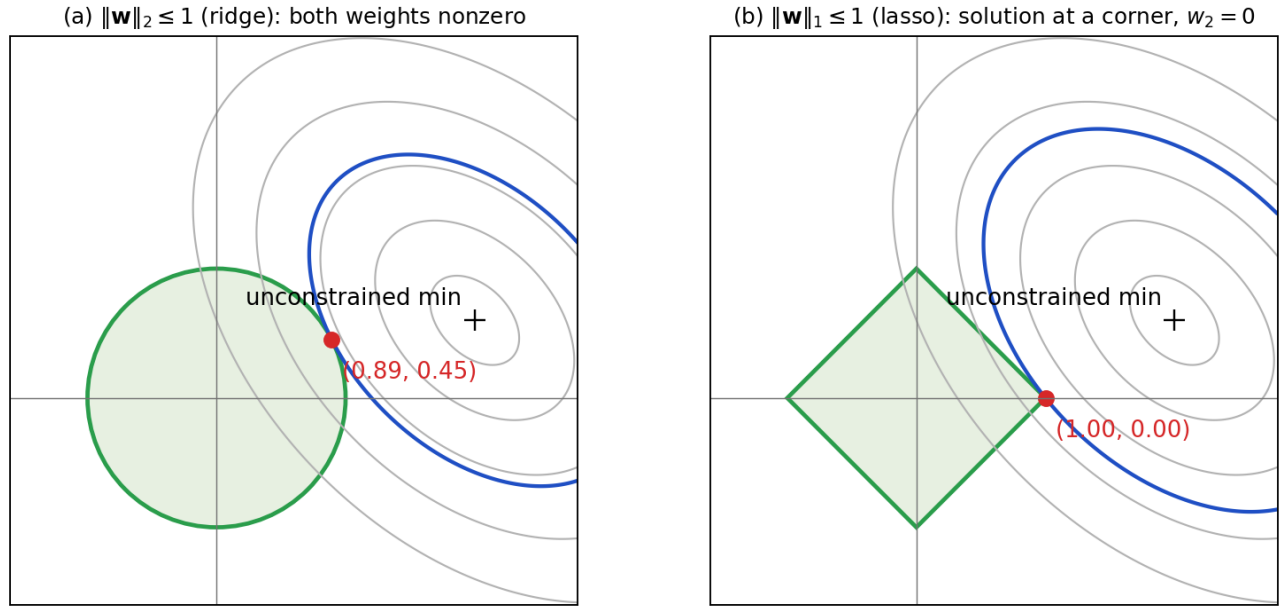


Figure 15: The same quadratic loss (contours) minimized over the L2 ball and over the L1 diamond of radius 1. The first contour to touch the ball touches it at a generic point; the first contour to touch the diamond touches it at a corner, so one weight is exactly zero.

- **Unconstrained:** bisection (1-D, linear rate), Newton (quadratic rate, needs the Hessian), gradient descent (cheap, zig-zags on elongated valleys), conjugate gradient, coordinate descent.
- **Equality constraints:** at the optimum the level curve of f is tangent to the constraint, $\nabla f = -\sum_j \beta_j \nabla h_j$; solve $\nabla L = 0$ for the candidates.
- **Inequality constraints:** each is active ($g_i = 0$, $\alpha_i \geq 0$) or inactive ($g_i < 0$, $\alpha_i = 0$); the **KKT conditions** are stationarity, primal feasibility, dual feasibility, complementary slackness.
- **Duality:** $\theta(\alpha, \beta) = \min_{\mathbf{w}} L$ is always concave and always a lower bound ($d^* \leq p^*$). For convex problems with a strictly feasible point (Slater) the gap is zero, and optimal $\Leftrightarrow$ KKT $\Leftrightarrow$ saddle point.
- **Algorithms:** projected gradient, penalty and log-barrier (interior point, with the duality-gap bound k/t), dual ascent.
- **Regularization** is a constraint in Lagrangian form; the L1 ball's corners are why lasso is sparse.