Fundamentals of Analytical Optimization

Karl Stratos
me@karlstratos.com

Last updated: July, 2025

Abstract

This note derives fundamental tools for analytically minimizing $f(x) \in \mathbb{R}$ over a feasible set $\mathcal{P}$ defined by inequality and equality constraints. Under certain conditions, a local optimum is a KKT point: a stationary point of the Lagrangian where the gradient of f balances the gradients of the active constraints (18). The Lagrangian is naturally motivated as an adversarial enforcement of the constraints (31). Switching the order of the players leads to the simple yet elegant theory of duality, which allows us to guarantee the existence and optimality of a KKT point for convex problems holding strong duality (Section 6). The convex theory extends to generalized inequalities and can handle structured constraints such as positive-definiteness of a matrix (Section 7). More generally, we may examine the geometry of feasible directions (Figure 1). Under mild requirements, most usefully LICQ (i.e., active constraints have linearly independent gradients), feasible directions form a linear cone and a theorem of the alternative forces every local optimum to satisfy the KKT conditions. Thus convex or not, one can enumerate all KKT points to obtain 100% recall on globally optimal solutions. This allows us to easily analyze nonconvex problems. For instance, we discover that a quadratic objective with mixed eigenvalues may be full of saddle points (Lemma 4.4) but exactly solvable—one of the small miracles in optimization.

Cheatsheet for practitioners

- If f and the inequality constraints are convex; the equality constraints are affine; and either
  - f and the inequality constraints are also affine (linear program), or
  - The interior of $\mathcal{P}$ is not empty (aka. Slater's condition);

  then strong duality holds; any KKT point is a global minimum.
- Otherwise, first make sure that LICQ holds at all $x \in \mathcal{P}$ so that all local minima are KKT points. Then find all KKT points $x_1 \dots x_n \in \mathcal{P}$ and select $x^\star = \arg\min_i \; f(x_i)$.

Contents

1 Feasible Directions **3**
1.1 Linearized Tangent Cone . . . 3
1.2 Secant Cone . . . 4

2 Minimization **5**
2.1 Local Minimum . . . 5
2.2 Optimality Test . . . 5
2.2.1 Convex loss . . . 7
2.3 Local Maximum and Saddle Point . . . 7

3 Lagrangian **8**
3.1 KKT Conditions: Necessity . . . 8
3.2 KKT Conditions: Sufficiency . . . 9
3.2.1 Constraint qualifications (CQs) . . . 10
3.2.2 Convex case . . . 10

4 Ball Constraint **10**
4.1 Linear Objective . . . 11
4.2 Quadratic Objective . . . 11
4.3 Convex Objective . . . 12

5 Other Applications **12**
5.1 Scaling Law 12

6 Duality **13**
6.1 KKT Conditions and Strong Duality 14
6.2 SVM Dual 15
6.3 Linear Programs 16

7 Matrix Extensions **17**
7.1 Semidefinite Programs (SDP) 18
7.2 AdaGrad Lemma 19

A Illustrations **20**
A.1 Linearized Feasible Directions (Section 1.1) 20
A.2 KKT Conditions: Necessity (Section 3.1) 20

B Approximation Error of a Two-Layer Regressor **20**

C Chinchilla Approaches **21**
C.1 Approach 1 22
C.2 Approach 2 22
C.3 Approach 3 23
C.3.1 Data-constrained extension 24

D Conic Optimization **25**
D.1 Convex Cones and Generalized Inequality 25
D.2 Generalized KKT Conditions 26
D.3 Generalized Duality 26

E Proofs and Lemmas **27**

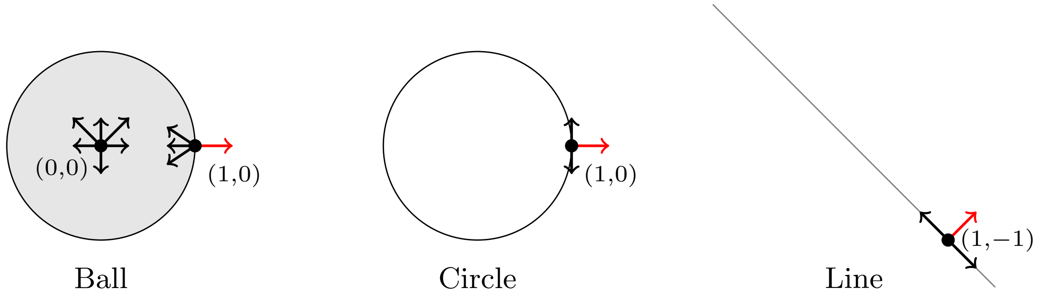


Figure 1: *Linearizing* the tangent cone allows for a simple geometric view in which feasible directions must move "away" from the active constraint gradients (annotation in Appendix A.1).

1 Feasible Directions

Any hypothesis space $\mathcal{P} \subseteq \mathbb{R}^d$ can be expressed with m inequality and r equality constraints:

$$\mathcal{P} = \left\{x \in \mathbb{R}^d : \; h(x) \leq 0_m, \; l(x) = 0_r\right\} \tag{1}$$

We assume that $\mathcal{P}$ is nonempty, and that $h_i(x) \in \mathbb{R}$ and $l_j(x) \in \mathbb{R}$ are continuously differentiable. $\mathcal{P}$ is not necessarily closed, e.g., $\mathcal{P} = \left\{x \in \mathbb{R} : \frac{1}{x} \geq 0\right\} = (0, \infty)$. The set of feasible directions at $x \in \mathcal{P}$ is given by

$$T(x) = \left\{t \in \mathbb{R}^d : \; \exists \epsilon > 0 \text{ such that } x + \eta t \in \mathcal{P} \;\; \forall \eta \in (0, \epsilon)\right\} \tag{2}$$

(2) is also called the **tangent cone**.[1] Note that $T(x)$ is nontrivial only at boundary points where some contraint is "active" since $T(x) = \mathbb{R}^d$ at interior points. We will write

$$I(x) = \{i : h_i(x) = 0\} \subseteq \{1 \ldots m\}$$

to denote the active inequality constraints at $x \in \mathcal{P}$.

1.1 Linearized Tangent Cone

We may define a first-order approximation $T_{\text{linear}}(x) \approx T(x)$, aka. the **linearized tangent cone** (Figure 1):

$$T_{\text{linear}}(x) := \left\{t \in \mathbb{R}^d : \; \nabla h_i(x)^\top t \leq 0 \quad \forall i \in I(x) \;\; \bigwedge \;\; \nabla l_j(x)^\top t = 0 \quad \forall j\right\} \tag{3}$$

(3) is a necessary condition for feasible directions since the first-order term dominates lower-order terms. If the constraints happen to be locally so simple that they can be completely specified by the first-order information, (3) is also a sufficient condition. All omitted proofs are in Appendix E unless said otherwise.

Lemma 1.1. At any $x \in \mathcal{P}$, we have $T(x) \subseteq T_{\text{linear}}(x)$.

Lemma 1.2. Let $x \in \mathcal{P}$ and suppose that in a neighborhood of x: (1) h_i is concave for all $i \in I(x)$ and (2) l_j is affine for all j. Then $T_{\text{linear}}(x) \subseteq T(x)$.

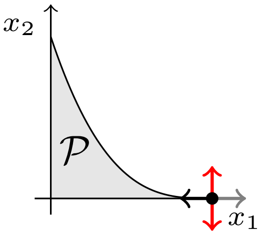
However, $T_{\text{linear}}(x)$ may contain spurious directions if the first-order information is not enough. The quickest way to see this is when all active constraints have zero gradients. Then $T_{\text{linear}}(x) = \mathbb{R}^d$, but a nonlinear constraint may still influence directions. Consider the example

$$\mathcal{P} = \left\{x \in \mathbb{R} : \; x^3 \leq 0\right\} = (-\infty, 0]$$

[1] "Tangent" because it characterizes the surrounding topology at x, and "cone" because if $t \in T(x)$ then $\alpha t \in T(x)$ for any $\alpha > 0$.

At $x = 0$ (i.e., the saddle point of the inequality constraint $h(x) = x^3$), it is clear that $T(0) = (-\infty, 0]$ by the original definition (2). But since $h'(0) = 0$, we have $T_{\text{linear}}(0) = (-\infty, \infty)$.[2] Thus we at least need active constraint gradients to be nonzero at boundary points. However, this may not be enough if there are multiple active constraints. Consider the example

$$\mathcal{P} = \{x \in \mathbb{R}^2 : \ x_1 \geq 0, \ x_2 \geq 0, \ x_2 - (1 - x_1)^3 \leq 0\}$$



At the boundary point $x = (1, 0)$, it is clear that $T(1, 0) = \{(x_1, 0) : x_1 \leq 0\}$. The two inequality constraints $h_2(x) = -x_2$ and $h_3(x) = x_2 - (1 - x_1)^3$ are active with gradients $\nabla h_2(x) = (0, -1)$ and $\nabla h_3(x) = (3(1 - x_1)^2, 1)$. Since $\nabla h_2(1, 0) = (0, -1)$ and $\nabla h_3(1, 0) = (0, 1)$, we have the wrong tangent cone $T_{\text{linear}}(1, 0) = \{(x_1, 0) : x_1 \in \mathbb{R}\}$. Intuitively, the gradients fail to characterize the interaction between the active constraints (despite being nonzero) because their directions are redundant. This motivates LICQ.

Definition 1.1. We say **linear independence constraint qualification (LICQ)** holds at $x \in \mathcal{P}$ if all active constraint gradients are linearly independent.

Note that LICQ requires the active constraint gradients to be nonzero. It also implies that the number of active constraints is at most d. This turns out to be a sufficient condition for every $t \in T_{\text{linear}}(x)$ to be a genuine feasible direction.

Lemma 1.3. If LICQ holds at $x \in \mathcal{P}$, then $T_{\text{linear}}(x) \subseteq T(x)$.

The proof is fairly sophisticated and relies on the implicit function theorem. It is given in Appendix E.

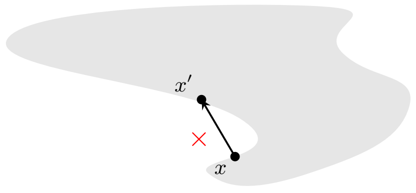
Summary. The linearized tangent cone overestimates the tangent cone. It is exact if active constraints are simple (concave for inequality, affine for equality) or have linearly independent gradients.

1.2 Secant Cone

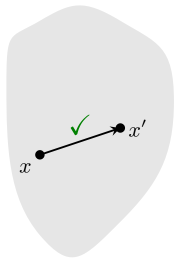
We may define the **secant cone** as the directions from $x \in \mathcal{P}$ to all feasible points, more formally

$$T_{\text{secant}}(x) = \left\{\alpha(x' - x) \in \mathbb{R}^d : \ x' \in \mathcal{P}, \ \alpha \in (0, \infty)\right\} \tag{4}$$

("secant" because two points cut through $\mathcal{P}$). It is intuitively clear that $T(x) \subseteq T_{\text{secant}}(x)$ with equality if $\mathcal{P}$ is convex. See the proof by picture (or read the footnote).[3]



Nonconvex $\mathcal{P}$



Convex $\mathcal{P}$

[2] As an exercise, we can "fix" this example by using higher-order information. Since $h'(0) = h''(0) = 0$ and $h'''(0) = 6$, we have $h(0 + \eta t) = h(0) + \frac{\eta^3 t^3}{6} h'''(0) + o(\eta^3)$ and obtain a correct third-order characterization $T_{\text{cubic}}(0) = \left\{t \in \mathbb{R} : t^3 \leq 0\right\} = T(0)$.

[3] If $t \in T(x)$, $x' = x + \eta t \in \mathcal{P}$ for some small enough $\eta > 0$, thus $t = \frac{1}{\eta}(x' - x) \in T_{\text{secant}}(x)$. Conversely, if $t = \alpha(x' - x) \in T_{\text{secant}}(x)$, then $x + \eta t = (1 - \eta\alpha)x + \eta\alpha x'$ is the step result. If $\mathcal{P}$ is convex, this is in $\mathcal{P}$ for all $\eta \in (0, \frac{1}{\alpha})$.

Summary. The secant cone overestimates the tangent cone. It is exact if the feasible set is convex.

2 Minimization

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable on $\mathcal{P}$ (1). We think of f as "loss" and solve

$$f^\star = \min_{x \in \mathcal{P}} \; f(x) \tag{5}$$

A feasible point $x^\star \in \mathcal{P}$ achieving $f^\star = f(x^\star)$ is called a **global minimum** of (5). Equivalently,

$$f(x^\star) \leq f(x) \qquad \forall x \in \mathcal{P} \tag{6}$$

$x^\star$ does not necessarily exist even if f is bounded below. For instance, $x^2 \geq 0$ is not minimized by any $x \in (0, \infty)$ (we can encode $(0, \infty)$ by the inequality $h(x) = -\frac{1}{x} \leq 0$). To handle this in the literature, authors often define

$$f^\star = \inf_{x \in \mathcal{P}} \; f(x)$$

which always exists under mild assumptions ($\mathcal{P}$ is nonempty and f is bounded below on $\mathcal{P}$, a consequence of the completeness of the real numbers), then say no $x \in \mathcal{P}$ may achieve $f^\star$. However, since we are almost always interested in actually finding a minimum, we will avoid this treatment and simply say $x^\star$ does not exist.[4]

2.1 Local Minimum

We say $x \in \mathcal{P}$ is a (one-sided) **local minimum** of f if there is some $\epsilon > 0$ such that for all $\eta \in (0, \epsilon)$

$$f(x + \eta t) \geq f(x) \qquad \forall t \in T(x) \tag{7}$$

That is, in every feasible direction f is either *locally* constant or *locally* increasing. If f is locally increasing in every $t \in T(x)$, we call x a **strict local minimum**. It is clear from (6) that a global minimum is a local minimum. We think of $f(x + \eta t)$ as applying progressively small local corrections on $f(x)$:

$$f(x + \eta t) = f(x) + \sum_{k=1}^{\infty} \frac{\eta^k}{k!} \nabla^k f(x).\mathbf{contract}(t) \tag{8}$$

where $\nabla^k f(x).\mathbf{contract}(t) \in \mathbb{R}$ is a k-th order polynomial of $t \in \mathbb{R}^d$ obtained by contracting the tensor $\nabla^k f(x)$ (e.g., $\nabla f(x)^\top t$ for $k = 1$). We write $R_{k+1}(x + \eta t) = \sum_{i > k} \frac{\eta^i}{i!} \nabla^i f(x).\mathbf{contract}(t)$ to denote the remainder after k expansions.

Fact 2.1. If f is locally constant in $t \in T(x)$, then $\nabla^k f(x).\mathbf{contract}(t) = 0$ for all $k \in \mathbb{N}$.

Fact 2.2. If f is locally increasing in $t \in T(x)$, there is some $k \in \mathbb{N}$ with $\nabla^k f(x).\mathbf{contract}(t) > 0$ such that for all sufficiently small $\eta > 0$

$$f(x + \eta t) = f(x) + \frac{\eta^k}{k!} \nabla^k f(x).\mathbf{contract}(t) + R_{k+1}(x + \eta t) \tag{9}$$

2.2 Optimality Test

We obtain a K-th order polynomial approximation of $f(x + \eta t)$ around x by cutting off (8) after the first K expansions:

$$f^{(K)}(x + \eta t) = f(x) + \sum_{k=1}^{K} \frac{\eta^k}{k!} \nabla^k f(x).\mathbf{contract}(t) \tag{10}$$

[4] To guarantee the existence of $x^\star$, we need stronger assumptions (e.g., if $\mathcal{P}$ is compact and f is continuous, we can invoke the extreme value theorem).

```
IsLocalMin
Input: $x \in \mathcal{P}$, $K \in \mathbb{N}$
Output: YES if $x$ is a strict local minimum of $f$, NO if $x$ is not a local minimum of $f$, MAYBE otherwise
  • Initialize $T'(x) \leftarrow T(x)$. For $k = 1 \dots K$
      – If $\nabla^k f(x).\mathbf{contract}(t) < 0$ for some $t \in T'(x)$, return NO.
      – If $\nabla^k f(x).\mathbf{contract}(t) > 0$ for all $t \in T'(x)$, return YES.
      – Update $T'(x) \leftarrow \big\{t \in T'(x) :\ \nabla^k f(x).\mathbf{contract}(t) = 0\big\}$.
  • Return MAYBE.
```

Figure 2: An iterative algorithm for analyzing the local minimality of a feasible point x up to the K-th order test. With $T(x) = \mathbb{R}^d$ and $K = 2$, the algorithm reduces to the familiar unconstrained optimality test where we first check if x is stationary, then check the eigenvalues of $\nabla^2 f(x)$ (if all positive, a strict local minimum; if all nonnegative, possibly a local minimum; if some negative, not a local minimum).

Lemma 2.3. A local minimum $x \in \mathcal{P}$ of f is a local minimum of $f^{(K)}$. More specifically,

- If f is locally constant in $t \in T(x)$, then $f^{(K)}$ is constant in t.
- If f is locally increasing in $t \in T(x)$, then $f^{(K)}$ is locally increasing in t.

The converse is not true because Fact 2.1 can only be used one way. If $f^{(K)}$ is locally constant in $t \in T(x)$, it could simply be that it does not have enough information rather than f being locally constant in t. If f "reveals" its true local behavior in a higher order term as *decreasing*, not constant or increasing, x is not a local minimum.

Example 2.1. Let $f(x) = x^3$ over $x \in \mathbb{R}$. We have $f^{(2)}(x+a) = x^3+3ax^2+3a^2x$ with the remainder $R_3(x+a) = a^3$. At $x = 0$, all expansion terms up to order $K = 2$ vanish, so $f^{(2)}(0+a) = f^{(2)}(0) = 0$. In particular, $f^{(2)}$ is constant in direction $t = -1$. However, $f(0 + \eta t) = -\eta^3 < 0 = f(0)$ for any $\eta > 0$, thus f is decreasing in t.

In contrast, Fact 2.2 *can* be used the other way and we have the following result.

Lemma 2.4. If $f^{(K)}$ is locally increasing in $t \in T(x)$, then f is also locally increasing in t.

Corollary 2.5. If x is a *strict* local minimum of $f^{(K)}$, it is also a strict local minimum of f.

By Lemma 2.3, we can rule out $x \in \mathcal{P}$ from being a local minimum of f unless x is a local minimum of $f^{(K)}$. Checking the latter amounts to showing

$$\sum_{k=1}^{K} \frac{\eta^k}{k!} \nabla^k f(x).\mathbf{contract}(t) \geq 0 \qquad \forall t \in T(x) \tag{11}$$

for all small enough $\eta > 0$. We can eliminate the dependence on η by progressively increasing K. When $K = 1$, (11) simplifies to

$$\nabla f(x)^\top t \geq 0 \qquad \forall t \in T(x) \qquad\qquad (\textbf{first-order test}) \tag{12}$$

If $T(x) = \mathbb{R}^d$, it is equivalent to $\nabla f(x) = 0_d$ (i.e., x is **stationary** or **critical**). Since $T(x) \subseteq T_{\text{secant}}(x)$ (see (4)), the following condition

$$\nabla f(x)^\top (x' - x) \geq 0 \qquad \forall x' \in \mathcal{P} \tag{13}$$

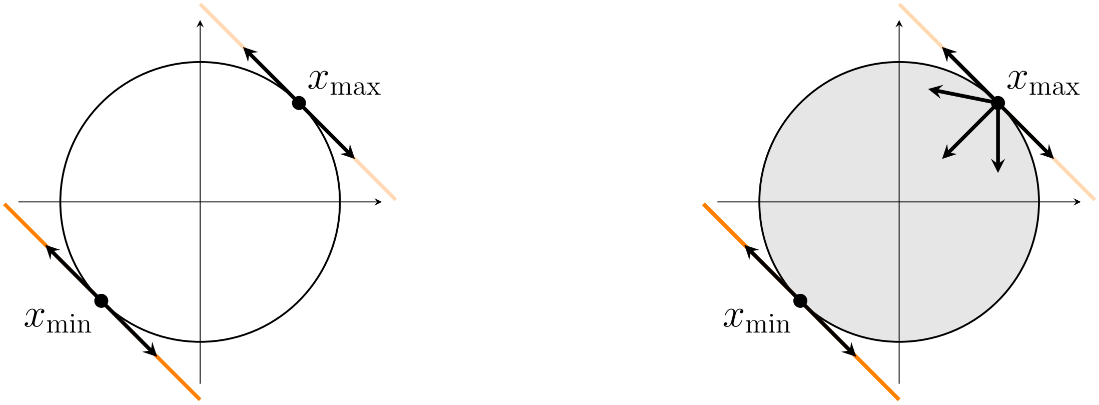


Figure 3: Minimizing $f(x) = x_1 + x_2$ (not strictly convex). On the nonconvex circle (left) it has a non-minimum $x_{\max}$ passing the first-order test since the loss stays constant in feasible directions. Over the convex ball (right), it fails the first-order test since it has a loss-reducing direction toward $x_{\min}$.

implies (12). They are equivalent if $\mathcal{P}$ is convex (Section 1.2). If x fails the first-order test, it is definitely not a local minimum. If it passes, we only need to check $t \in T(x)$ such that $\nabla f(x)^\top t = 0$ since f is locally increasing in t if $\nabla f(x)^\top t > 0$ (Lemma 2.4). For such directions, we have $f^{(2)}(x) = f(x) + \frac{\eta^2}{2} t^\top \nabla^2 f(x) t$, thus (11) simplifies to (with $K = 2$)

$$t^\top \nabla^2 f(x) t \geq 0 \quad \forall t \in T(x) : \; \nabla f(x)^\top t = 0 \qquad \textbf{(second-order test)} \tag{14}$$

If $T(x) = \mathbb{R}^d$, combined with the first-order test which ensures $\nabla f(x) = 0_d$, it is equivalent to $\nabla^2 f(x) \succeq 0$. We continue until no feasible direction remains whose lower-order contractions vanish. At this point, every feasible direction already yields a strictly positive leading derivative, thus x is a strict local minimum of $f^{(K)}$ and must be a strict local minimum of f. The verification algorithm is given in Figure 2.

2.2.1 Convex loss

If the loss f is convex and $x \in \mathcal{P}$ passes the first-order test (i.e., $\nabla f(x)^\top t \geq 0$ for all $t \in T(x)$), by the convexity of f we have for all $\eta > 0$

$$f(x + \eta t) \geq f(x) + \eta \nabla f(x)^\top t \geq f(x) \qquad \forall t \in T(x) \tag{15}$$

If f is *strictly* convex, the first inequality is strict and x is indeed a strict local minimum. Otherwise, it is possible that f is just constant in all feasible directions $t \in T(x)$ unless the feasible set is also convex.

Lemma 2.6. If both f and $\mathcal{P}$ are convex, $x \in \mathcal{P}$ is a global minimum iff it passes the first-order test.

Proof. One direction is given by the necessity of first-order optimality (i.e., if x is a global minimum, it has to pass the first-order test). For the other direction, let $x' \in \mathcal{P}$ be a different point with $f(x') < f(x)$ if there is any (otherwise we are done). We will show that there is a decreasing feasible direction at x', thus it fails the first-order test. Specifically, the direction is $t = x - x'$. To see feasibility, we note $x' + \eta t = (1-\eta)x' + \eta x \in \mathcal{P}$ for all $\eta \in [0,1]$ by the convexity of $\mathcal{P}$. To see decreasing, we note

$$f(x' + \eta t) = f((1-\eta)x' + \eta x) \leq (1-\eta) f(x') + \eta f(x) < f(x)$$

for all $\eta \in (0,1]$ by the convexity of f. □

See Figure 3 for an illustration. Since (13) and (12) are equivalent when $\mathcal{P}$ is convex, the lemma is often stated as:

$$x^\star = \arg\min_{x \in \mathcal{P}} f(x) \qquad \overset{(f,\mathcal{P}\text{ convex})}{\Leftrightarrow} \qquad \nabla f(x^\star)^\top (x - x^\star) \geq 0 \quad \forall x \in \mathcal{P} \tag{16}$$

2.3 Local Maximum and Saddle Point

We may define a **local maximum** analogously. A point $x \in \mathcal{P}$ that has both increasing and decreasing feasible directions $t, t' \in T(x)$ is called a **saddle point**. In particular, if x is stationary and the Hessian has both postive and negative eigenvalues, it is a saddle point. The different types of stationary points are illustrated in Figure 4.

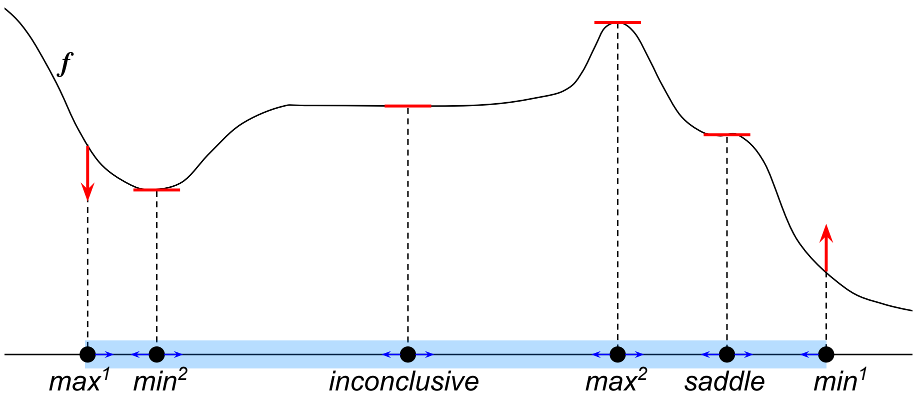


Figure 4: Optimality of f on various points over $\mathcal{P} \in \mathbb{R}$ (the blue interval). The superscript indicates if the minimum or maximum can be identified by the first-order or the second-order test.

3 Lagrangian

Given a loss function $f : \mathbb{R}^d \to \mathbb{R}$ over $\mathcal{P} = \left\{x \in \mathbb{R}^d : \ h(x) \leq 0_m, \ l(x) = 0_r\right\}$, we first focus on identifying $x \in \mathcal{P}$ that passes the first-order test (12)

$$\nabla f(x)^\top t \geq 0 \qquad \forall t \in T(x)$$

upon which we may follow up with additional analysis (e.g., second-order test). Unless $T(x) = \mathbb{R}^d$, we cannot just hunt for $x \in \mathcal{P}$ that is stationary in f since it may pass the test with $\nabla f(x) \neq 0_d$. Instead, we define a helper function $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ called the **Lagrangian**

$$L(x, \rho, \lambda) := f(x) + \rho^\top h(x) + \lambda^\top l(x) \tag{17}$$

Here, $\rho \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}^r$ are called the **Lagrangian multipliers** corresponding to the m inequality and r equality constraints.

3.1 KKT Conditions: Necessity

Definition 3.1. We say $x \in \mathbb{R}^d$ satisfies the **KKT conditions** if we can find $\rho \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}^r$ such that

1. **Primal feasibility**: $x \in \mathcal{P}$
2. **Dual feasibility**: $\rho \geq 0_m$
3. **Complementary slackness**: $\rho_i = 0$ whenever $h_i(x) < 0$
4. **Stationarity**: $\nabla_x L(x, \rho, \lambda) = 0_d$

Combined with primal feasibility, complementary slackness can be written in the product form: $\rho_i h_i(x) = 0$ for each $i = 1 \ldots m$. Then the stationarity condition asserts that

$$\nabla f(x) = -\sum_{i \in I(x)} \rho_i \nabla h_i(x) - \sum_{j=1}^{r} \lambda_j \nabla l_j(x) \tag{18}$$

where $I(x) = \{i : h_i(x) = 0\} \subseteq \{1 \ldots m\}$. Geometrically, $\nabla f(x) \in \mathbb{R}^d$ must be "cancelled out" by some restricted linear combination of $\nabla h_i(x), \nabla l_j(x) \in \mathbb{R}^d$ where h_i, l_j are active constraints at x. Another useful way to summarize the KKT conditions is that $\nabla f(x) \in \mathbb{R}^d$ must lie inside the cone spanned by active constraint gradients under dual feasibility ("KKT cone")

$$K(x) = \{-\nabla h(x)\rho - \nabla l(x)\lambda : \lambda \in \mathbb{R}^r, \ \rho \geq 0_m \text{ with } \rho_i = 0 \text{ for } i \notin I(x)\} \subseteq \mathbb{R}^d \tag{19}$$

where $\nabla h(x) \in \mathbb{R}^{d \times m}$ and $\nabla l(x) \in \mathbb{R}^{d \times r}$ are Jacobians.

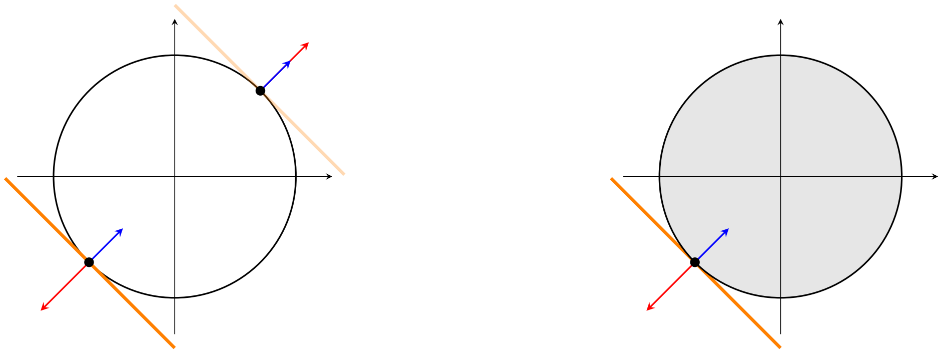


Figure 5: We can visually verify Lemma 3.1 on the example in Figure 3 (annotation in Appendix A.2). The circle $\mathcal{P}$ corresponds to an equality constraint $l(x) = 0$ and has two KKT points $x \in \mathcal{P}$ where $\nabla f(x) = \pm\frac{\sqrt{2}}{2}\nabla l(x)$. The ball $\mathcal{P}$ corresponds to an inequality constraint $h(x) \leq 0$ and has one KKT point $x \in \mathcal{P}$ where $\nabla f(x) = -\frac{\sqrt{2}}{2}\nabla h(x)$. Note that $\nabla f(x)$ and $\nabla h(x)$ point in the opposite opposite directions by dual feasibility.

Lemma 3.1. If $x \in \mathbb{R}^d$ satisfies the KKT conditions, it passes the first-order test.

Proof. Pick any $t \in T(x)$. Since $T(x) \subseteq T_{\text{linear}}(x)$, we have $\nabla h_i(x)^\top t \leq 0$ for all $i \in I(x)$ and $\nabla l_j(x)^\top t = 0$ for all j (see (3)). From (18), we have

$$\nabla f(x)^\top t = -\sum_{i \in I(x)} \rho_i \nabla h_i(x)^\top t - \sum_{j=1}^{r} \lambda_j \nabla l_j(x)^\top t \geq 0$$

□

Thus any point that passes the first-order test necessarily satisfies the KKT conditions. See Figure 5 for an illustration.

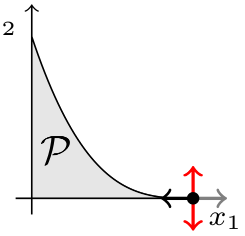
3.2 KKT Conditions: Sufficiency

The proof of Lemma 3.1 relies on the fact that $T(x) \subseteq T_{\text{linear}}(x)$ to linearize a feasible direction (Section 1.1). Since the converse does not hold, we should not be surprised that $x \in \mathcal{P}$ may pass the first-order test but fail the KKT conditions if $T(x) \neq T_{\text{linear}}(x)$. For instance, recall the example

$$\mathcal{P} = \left\{x \in \mathbb{R} : \; x^3 \leq 0\right\} = (-\infty, 0]$$

where $T(0) = (-\infty, 0]$ and $T_{\text{linear}}(0) = (-\infty, \infty)$. If we minimize $f(x) = -x$ over $\mathcal{P}$, clearly $x = 0$ is a local minimum and thus passes the first-order test. But since $\nabla f(0) = -1$ is not in the span of the active constraint gradient $\nabla h(0) = 0$, it cannot satisfy the KKT conditions. Similarly, recall the example

$$\mathcal{P} = \left\{x \in \mathbb{R}^2 : \; x_1 \geq 0, \; x_2 \geq 0, \; x_2 - (1 - x_1)^3 \leq 0\right\}$$



where $T(1,0) = \{(x_1, 0) : x_1 \leq 0\}$ and $T_{\text{linear}}(1,0) = \{(x_1, 0) : x_1 \in \mathbb{R}\}$. If we minimize $f(x) = -x_1$ over $\mathcal{P}$, $x = (1,0)$ is a local minimum and thus passes the first-order test. But since $\nabla f(1,0) = (-1, 0)$ is not in the span of the two active constraint gradients $\nabla h_2(1,0) = (0,-1)$ and $\nabla h_3(1,0) = (0,1)$, it cannot satisfy the KKT conditions.

3.2.1 Constraint qualifications (CQs)

Given the above cases of degeneracy, the following result is natural.

Lemma 3.2. If $x \in \mathcal{P}$ is a local minimum of f and $T(x) = T_{\text{linear}}(x)$, it satisfies the KKT conditions.

Proof. It is sufficient to show that $\nabla f(x)$ is inside the KKT cone (19). A classical theorem of the alternative gives exactly this result (see Nocedal and Wright (1999) for a proof).

Farkas' lemma Let $K = \{B\rho + C\lambda : \rho \geq 0_m, \lambda \in \mathbb{R}^r\} \subseteq \mathbb{R}^d$ be a cone parameterized by any $B \in \mathbb{R}^{d \times m}$ and $C \in \mathbb{R}^{d \times r}$. Given any $g \in \mathbb{R}^d$, either $g \in K$ or there is some seperating hyperplane $t \in \mathbb{R}^d$ such that (i) $g^\top t < 0$, (ii) $B^\top t \geq 0_m$, and (iii) $C^\top d = 0_r$.

Since x passes the first-order test and $T(x) = T_{\text{linear}}(x)$, every $t \in \mathbb{R}^d$ satisfying $\nabla h_{\text{active}}(x)^\top t \leq 0_{m_x}$ and $\nabla l(x)^\top t = 0_r$ satisfies $\nabla f(x)^\top t \geq 0$. By Farkas' lemma,

$$\nabla f(x) \in \{-\nabla h_{\text{active}}(x)\rho_{\text{active}} - \nabla l(x)\lambda : \rho_{\text{active}} \geq 0_m,\ \lambda \in \mathbb{R}^r\}$$

Pick any associated $\rho_{\text{active}}, \lambda$ and define $\rho \in \mathbb{R}^m$ by $\rho_i = \rho_{\text{active},i}$ if $i \in I(x)$ and $\rho_i = 0$ otherwise. Then $\nabla f(x) = -\nabla h(x)\rho - \nabla l(x)\lambda \in K(x)$. □

The following corollaries are immediate from Lemma 1.2 and 1.3.

Corollary 3.3. If $x \in \mathcal{P}$ is a local minimum of f and LICQ holds at x, it satisfies the KKT conditions.

Corollary 3.4. If $x \in \mathcal{P}$ is a local minimum of f and h_i is locally concave for all $i \in I(x)$ and l_j is locally affine for all j, it satisfies the KKT conditions.

Summary. KKT⇒first-order test requires no CQs. Locally optimal⇒KKT requires CQs. Beware that even with CQs, not all KKT points are necessarily locally optimal; they are simply guaranteed to contain all locally optimal solutions, and may contain non-solutions that pass the first-order test (e.g., saddle points).

3.2.2 Convex case

The following constraint qualification is central in convex analysis. While it is possible to prove this result without relying on duality, the argument is simpler with it, so we defer the proof to Section 6.

Lemma 3.5. Suppose that $f, h_1 \ldots h_m, l_1 \ldots l_r$ are all affine, *or*

1. f and $h_1 \ldots h_m$ are convex; $l_1 \ldots l_r$ are affine, and
2. (**Slater's condition**) There exists a strictly feasible point, namely $x \in \mathbb{R}^d$ such that $h(x) < 0_m$ and $l(x) = 0_r$,

Then $x \in \mathcal{P}$ is a global minimum of f iff it satisfies the KKT conditions.

4 Ball Constraint

A common feasible set is a "ball" $\mathcal{P} = \{x : h(x) \leq 0\}$ where

$$h(x) = ||x|| - C \tag{20}$$

$||\cdot|| : \mathbb{R}^d \to \mathbb{R}$ is some norm (thus convex) and $C > 0$ controls the radius. One class of norms is the "weighted" Euclidean norm

$$||x||_A = \sqrt{x^\top A x} \qquad (A \succ 0) \tag{21}$$

yielding a weighted Euclidean ball. Another class of norms is the l_p norm

$$||x||_p = (|x_1|^p + \cdots + |x_d|^p)^{1/p} \qquad (p \geq 1) \tag{22}$$

yielding an l_p ball.[5] Both classes include the standard Euclidean norm as a special case (with $A = I_d$ and $p = 2$).

4.1 Linear Objective

Pick any $A \succ 0$ and a nonzero $g \in \mathbb{R}^d$ and consider

$$f^\star = \min_{x \in \mathbb{R}^d:\, ||x||_A \leq C} g^\top x \tag{23}$$

Note that $g = \nabla f(x)$. See Figure 5 right for an illustration.

Lemma 4.1. (23) is uniquely minimized by $x^\star = -\eta A^{-1} g$ where $\eta = C / ||g||_{A^{-1}} > 0$. The minimum value is $f^\star = -C\, ||g||_{A^{-1}}$.

Example 4.1 (Steepest descent)**.** Let $l : \mathbb{R}^d \to \mathbb{R}$ be a loss function differentiable at $\theta \in \mathbb{R}^d$ (and not flat). The **steepest descent** step $\delta \in \mathbb{R}^d$ minimizes the linear approximation $l(\theta + \delta) \approx l(\theta) + \nabla l(\theta)^\top \delta$ around θ. Since this approximation is only locally accurate, we limit the size of δ by $||\delta||_A \leq 1$. By Lemma 4.1, we have

$$\delta^\star = -\frac{1}{||\nabla l(\theta)||_{A^{-1}}} A^{-1} \nabla l(\theta)$$

yielding the loss reduction $l(\theta + \delta) \approx l(\theta) - ||\nabla l(\theta)||^2_{A^{-1}}$. Gradient descent, Newton's method, and natural gradient are special cases of this solution with A equal to the identity, Hessian, and Fisher information matrix.

Example 4.2 (Dual norm)**.** For any norm $||\cdot|| : V \to \mathbb{R}$ on a vector space V, the **dual norm** $||\cdot||_* : V \to \mathbb{R}$ is defined as $||v||_* = \max_{w \in V:\, ||w|| \leq 1} v^\top w$. By Lemma 4.1, the dual norm of $||\cdot||_A : \mathbb{R}^d \to \mathbb{R}$ is

$$||v||_{A,*} = ||v||_{A^{-1}}$$

4.2 Quadratic Objective

Pick any nonzero symmetric matrix $H \in \mathbb{R}^{d \times d}$ (otherwise $x^\top H x$ may not be real) and consider[6]

$$f^\star = \min_{x \in \mathbb{R}^d:\, ||x||_2 \leq 1} x^\top H x \tag{24}$$

Note that $H = \nabla^2 f(x)$.

Lemma 4.2. If $H \succeq 0$, (24) is minimized by any $x^\star \in \mathrm{null}\,(H) \cap \{x : ||x||_2 \leq 1\}$ (in particular, uniquely by $x^\star = 0_d$ if $H \succ 0$). The minimum value is $f^\star = 0$.

Lemma 4.3. Suppose $H \not\succeq 0$. Let $v_1 \ldots v_d \in \mathbb{R}^d$ be orthonormal eigenvectors of H with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. Then (24) is minimized by $x^\star = v_d$. The minimum value is $f^\star = \lambda_d < 0$.

Lemma 4.4. Suppose $H \not\succeq 0$. Let $v_1 \ldots v_d \in \mathbb{R}^d$ be orthonormal eigenvectors of H with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{d-1} > \lambda_d$. Then $v_2 \ldots v_{d-1}$ are saddle points of (24).

Lemma 4.4 holds because we can take a step in v_1 to increase the loss and in v_d to decrease the loss from any of $v_2 \ldots v_{d-1}$. Remarkably, despite the saddle points, we can exactly solve (24) in $O(d^2)$ time (e.g., power iteration).

[5] (22) is convex for every $p \geq 1$ (including $p = \infty$, i.e., the uniform norm $||x||_\infty = \max_{i=1}^d |x_i|$).

[6] WLOG we fix the ball's shape $A = I_d$ and radius $C = 1$ (otherwise, use the substitution $y = C^{-1} A^{1/2} x$).

Example 4.3 (Top eigenvector)**.** While we use minimization in (24) to stay consistent with our loss-minimizing approach, the typical variational characterization of eigenvectors uses maximization. But we can easily cover this case by

$$f^\star = \max_{x \in \mathbb{R}^d:\ ||x||_2 \leq 1} x^\top H x = -\left(\min_{x \in \mathbb{R}^d:\ ||x||_2 \leq 1} x^\top (-H) x \right) \tag{25}$$

It follows that $f^\star = 0$ if $H \preceq 0$ (achieved by any vector of length ≤ 1 in the null space of H) and $f^\star = \lambda_1$ if H has a positive eigenvalue (achieved by v_1).

4.3 Convex Objective

Let $f : \mathbb{R}^d \to \mathbb{R}$ be any differentiable convex loss and $||\cdot|| : \mathbb{R}^d \to \mathbb{R}$ any norm. We deliberately keep the norm general (e.g., weighted Euclidian norm, l_p norm). The "hard" vs "soft" ball constraint is

$$f^\star_{\text{hard}} = \min_{x \in \mathbb{R}^d:\ ||x|| \leq C} f(x) \tag{26}$$

$$f^\star_{\text{soft}} = \min_{x \in \mathbb{R}^d} f(x) + D\,||x|| \tag{27}$$

Lemma 4.5. If $f^\star_{\text{hard}} = f(x^\star)$ in (26) is achieved "interestingly" by a boundary point $||x^\star|| = C$ with $\nabla f(x^\star) \neq 0_d$, then $x^\star$ is also a minimizer of (27) for some $D > 0$ (with $f^\star_{\text{soft}} = f^\star_{\text{hard}} + DC$).

Lemma 4.6. If $f^\star_{\text{soft}} = f(x^\star)$ in (27) is achieved "interestingly" by a nonzero $x^\star \in \mathbb{R}^d$, then $x^\star$ is also a (boundary-point) minimizer of (26) with $C = ||x^\star||$ (with $f^\star_{\text{soft}} = f^\star_{\text{hard}} + DC$).

Nonconvex objective. The equivalence between (26) and (27) only holds for convex functions, but practitioners often invoke this for nonconvex objectives as well (e.g., in deep learning). This is roughly justified if the region of interest is "convex enough", but in general there is no justification.

5 Other Applications

5.1 Scaling Law

The Chinchilla scaling law (Hoffmann *et al.*, 2022) predicts the expected loss $L(N, D) \in \mathbb{R}$ from the model and data sizes $N, D > 0$ as

$$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} \tag{28}$$

where $E, A, B, \alpha, \beta > 0$ are estimated from samples of (N, D, L) (Approach 3, Appendix C). A compute budge $C > 0$ yields the feasible range of N, D by

$$\mathcal{P} = \left\{ (N, D) \in \mathbb{R}^2 :\ 6ND = C \right\}$$

where we omit enforcing positivity since we can always filter out negative solutions later if we have any. LICQ holds on all of $\mathcal{P}$, thus any local optimum is a KKT point. The Lagrangian is

$$L(N, D, \lambda) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} + \lambda(6ND - C)$$

Stationarity $\nabla_{(N,D)} L(N, D, \lambda) = (0, 0)$ gives us

$$\begin{aligned} \frac{\partial L(N, D, \lambda)}{\partial N} &= -\frac{\alpha A}{N^{\alpha+1}} + 6\lambda D = 0 & \Leftrightarrow && \lambda = \frac{\alpha A}{6DN^{\alpha+1}} \\ \frac{\partial L(N, D, \lambda)}{\partial D} &= -\frac{\beta B}{D^{\beta+1}} + 6\lambda N = 0 & \Leftrightarrow && \lambda = \frac{\beta B}{6ND^{\beta+1}} \end{aligned}$$

from which we have $\frac{N^\alpha}{D^\beta} = \frac{\alpha A}{\beta B}$. Primal feasibility $(N, D) \in \mathcal{P}$ gives us $D = \frac{C}{6N}$ and $N = \frac{C}{6D}$. Solving the equations, we get

$$N = G\left(\frac{C}{6}\right)^b \qquad D = G^{-1}\left(\frac{C}{6}\right)^a \qquad G = \left(\frac{\alpha A}{\beta B}\right)^{\frac{1}{\alpha+\beta}} \qquad a = \frac{\alpha}{\alpha+\beta} \qquad b = \frac{\beta}{\alpha+\beta} \tag{29}$$

We see that $N, D > 0$, thus no need to worry about negative solutions.[7] Since the KKT point is unique, (29) must be the global minimum. It also establishes a relation between the optimal model and data sizes as

$$N = G\,(GD)^{\frac{\beta}{\alpha}} \qquad\qquad D = G^{-1}\left(G^{-1}N\right)^{\frac{\alpha}{\beta}} \tag{30}$$

Empirical fits show $\alpha \approx \beta$ (i.e., model and data should scale jointly). Practitioners often fix $\alpha = \beta$ and find that $D = G^{-2}N \approx 20N$.

6 Duality

It is easy to see that primal feasibility $\mathcal{P} = \left\{x \in \mathbb{R}^d : \; h(x) \leq 0_m, \; l(x) = 0_r\right\}$ can be enforced "softly" by

$$f^\star = \min_{x \in \mathcal{P}} \; f(x) = \min_{x \in \mathbb{R}^d} \; \max_{\substack{\rho \geq 0_m \\ \lambda \in \mathbb{R}^r}} \; L(x, \rho, \lambda) \tag{31}$$

where $L(x, \rho, \lambda) = f(x) + \sum_i \rho_i h_i(x) + \sum_j \lambda_j l_j(x)$ is the Lagrangian (17). This works because we propose $x \in \mathbb{R}^d$ first; the "enemy" can send the loss to infinity if x violates any constraint. But suppose we let him attack first. Specifically, the enemy *maximizes* the **dual function**

$$g(\rho, \lambda) := \min_{x \in \mathbb{R}^d} \; L(x, \rho, \lambda) \tag{32}$$

For simple problems, (32) has a closed-form solution and is potentially easier to analyze than the primal objective f. This yields the **dual problem**:

$$g^\star := \max_{\substack{\rho \geq 0_m \\ \lambda \in \mathbb{R}^r}} \; g(\rho, \lambda) = \max_{\substack{\rho \geq 0_m \\ \lambda \in \mathbb{R}^r}} \; \min_{x \in \mathbb{R}^d} \; L(x, \rho, \lambda) \tag{33}$$

Intuitively, the dual problem gives us an upper hand since we get a chance to counter the enemy's attack. Thus we always have **weak duality**:

$$g^\star \leq f^\star \tag{34}$$

(34) may not hold with equality and $f^\star - g^\star \geq 0$ is called the duality gap. If $f^\star = g^\star$, we say **strong duality** holds. There are known sufficient conditions for strong duality to hold: see the list on Wikipedia. We focus on two simplest conditions.

Fact 6.1. Strong duality holds if any of the following holds.

- (**Linear program**) $f, h_1 \ldots h_m, l_1 \ldots l_r$ are affine (Section 6.3).
- (**Convex program** + **Slater's condition**) $f, h_1 \ldots h_m$ are convex; $l_1 \ldots l_r$ are affine; there is a strictly feasible primal point $\bar{x}$, satisftying $h_i(\bar{x}) < 0$ for all i and $l_i(\bar{x}) = 0$ for all j.

Proof sketch. Let $\mathcal{C} = \{(f(x), h(x), l(x)) : x \in \mathbb{R}^d\} \subset \mathbb{R}^{1+m+r}$. For any $\alpha < f^\star$, $y = (\alpha, 0_m, 0_r)$ is outside $\mathcal{C}$.

- (Linear) $\mathcal{C}$ is an affine image of $\mathbb{R}^d$, so Farkas' lemma yields a separating hyperplane through y.
- (Convex+Slater) $\mathcal{C}$ is convex and contains $(f(\bar{x}), h(\bar{x}), l(\bar{x}))$ in its interior, so the supporting-hyperplane theorem provides the separation.

In either case we obtain a normal (θ, ρ, λ) with $\theta > 0$, $\rho \geq 0$, and $\theta\alpha \leq \theta f(x) + \rho^\top h(x) + \lambda^\top l(x)$ for all x. Rescaling by θ (set $\theta = 1$) gives $\alpha \leq g(\rho, \lambda) \leq g^\star \leq f^\star$. Letting $\alpha \uparrow f^\star$ forces $g^\star = f^\star$. □

[7] The paper estimates $A \approx B$ but $a = 0.46$ and $b = 0.54$. This marginally disagrees with Approach 1 and 2 where $a = b = 0.5$. The paper explains this is due to larger residuals on smaller models which are treated as outliers in Approach 3 with the Huber loss.

Example 6.1 (Revisiting (23))**.** Pick any $A \succ 0$, nonzero $g \in \mathbb{R}^d$, $C > 0$ and consider

$$f^\star = \min_{x \in \mathbb{R}^d:\ ||x||_A \leq C} g^\top x$$

We can easily derive the dual function and the dual optimum:

$$g(\rho) = \begin{cases} -\frac{1}{4\rho} g^\top A^{-1} g - \rho C^2 & \text{if } \rho > 0 \\ -\infty & \text{if } \rho = 0 \end{cases} \qquad\qquad g^\star = -C\,||g||_{A^{-1}}$$

achieved by $\rho^\star = ||g||_{A^{-1}} / (2C) > 0$. Since $||0_d||_A = 0 < C$, Slater's condition holds and we know off the bat that strong duality holds, thus $f^\star = g^\star = -C\,||g||_{A^{-1}}$ consistently with Lemma 4.1.

Example 6.2 (Exercise 5.21 of Boyd and Vandenberghe (2014))**.** Consider

$$f^\star = \min_{x,y>0:\ \frac{x^2}{y} \leq 0} e^{-x} = 1 \tag{35}$$

(achieved by $x^\star = 0$). Treating $y > 0$ as a domain restriction, we consider only the inequality constraint $h(x, y) = \frac{x^2}{y}$ which is convex over $x \in \mathbb{R}$ and $y > 0$. For all $\rho \geq 0$, the dual function is

$$g(\rho) = \inf_{x,y>0} e^{-x} + \rho \frac{x^2}{y} = \inf_x e^{-x} = 0$$

(the infimum is achieved in the limit $x, y \to \infty$ and not attained). Thus $g^\star = 0 < 1 = f^\star$ and strong duality fails, even though (35) is convex.

A central result is the following:

Lemma 6.2. *If $x \in \mathcal{P}$, $\rho \geq 0_m$, and $\lambda \in \mathbb{R}^r$ satisfy $f(x) = g(\rho, \lambda)$, then $f(x) = f^\star = g^\star = g(\rho, \lambda)$.*

Proof. The statement follows from weak duality $g(\rho, \lambda) \leq g^\star \leq f^\star \leq f(x)$ and the premise $f(x) = g(\rho, \lambda)$. □

Thus if we find any feasible points x, ρ, λ holding the "certificate of optimality" $f(x) = g(\rho, \lambda)$, strong duality holds—with them being solutions. This makes them a saddle point of the Lagrangian (assuming suitable constraint qualifications).[8]

6.1 KKT Conditions and Strong Duality

Lemma 6.3. *The KKT conditions (Definition 3.1) are necessary for strong duality.*

Proof. Suppose strong duality $f(x^\star) = f^\star = g^\star = g(\rho^\star, \lambda^\star)$ holds for some $x^\star \in \mathcal{P}$ and $\rho^\star \geq 0_m, \lambda^\star \in \mathbb{R}^r$. The primal and dual feasibility conditions are immediately satisfied. We observe

$$f(x^\star) = g(\rho^\star, \lambda^\star) := \min_{x \in \mathbb{R}^d} f(x) + \sum_{i=1}^m \rho_i^\star h_i(x) + \sum_{j=1}^r \lambda_j^\star l_j(x) \leq f(x^\star) + \sum_{i=1}^m \rho_i^\star h_i(x^\star) + \sum_{j=1}^r \lambda_j^\star l_j(x^\star) \leq f(x^\star)$$

The last inequality follows from the feasibility of $x^\star, \rho^\star, \lambda^\star$. Therefore the inequalities are equalities. To check stationarity, we write the first inequality as

$$L(x^\star, \rho^\star, \lambda^\star) = \min_{x \in \mathbb{R}^d} L(x, \rho^\star, \lambda^\star)$$

which shows that $x = x^\star$ is a stationary point of $L(x, \rho^\star, \lambda^\star)$, thus $0_d \in \partial_x L(x^\star, \rho^\star, \lambda^\star)$. To check complementary slackness, we write the last inequality as

$$\sum_{i=1}^m \rho_i^\star h_i(x^\star) = 0$$

Since $\rho_i^\star h_i(x^\star) \leq 0$ for all i, it must be that $\rho_i^\star h_i(x^\star) = 0$ for all i. □

[8] More specifically, since the KKT conditions must hold at x, ρ, λ, they are stationary points of the Lagrangian L. From here, we can increase L along x and decrease L along ρ, λ.

Lemma 6.4. If f and $h_1 \ldots h_m$ are convex, and $l_1 \ldots l_r$ are affine, then the KKT conditions are sufficient for strong duality.

Proof. By premise, the Lagrangian

$$L(x, \rho, \lambda) = f(x) + \sum_{i=1}^{m} \rho_i h_i(x) + \sum_{j=1}^{r} \lambda_j l_j(x) \tag{36}$$

is convex in $x \in \mathbb{R}^d$ for any $\rho \geq 0_m$, and $\lambda \in \mathbb{R}^r$ (the unrestricted sign of λ_j requires l_j to be affine so that $\lambda_j l_j(x)$ is affine and preserves convexity). Let $x^\star \in \mathcal{P}$ be a KKT point with Lagrangian multipliers $\rho^\star \geq 0_m, \lambda^\star \in \mathbb{R}^r$. By complementary slackness and primal feasibility, we have $L(x^\star, \rho^\star, \lambda^\star) = f(x^\star)$. By stationarity of $x^\star$ and the convexity of (36), $L(x^\star, \rho^\star, \lambda^\star) = g(\rho^\star, \lambda^\star)$. Since the two values match, $x^\star, \rho^\star, \lambda^\star$ hold the certificate of optimality (Lemma 6.2) and it must be that $f(x^\star) = f^\star = g^\star = g(\rho^\star, \lambda^\star)$. □

Lemma 6.4 does *not* guarantee that if the problem is convex (with affine equality), strong duality holds. This is because there may be no KKT point: (35) is an example. However, if we do have strong duality, the convexity assumption guarantees that an optimal solution is equivalent to a KKT point.

Lemma 6.5. Suppose strong duality holds: $f^\star = g^\star$. If f and $h_1 \ldots h_m$ are convex, $l_1 \ldots l_r$ are affine, then $x^\star \in \mathcal{P}$, $\rho^\star \geq 0_m$, and $\lambda^\star \in \mathbb{R}^r$ attain the optimum $f(x^\star) = f^\star = g^\star = g(\rho^\star, \lambda^\star)$ iff they satisfy the KKT conditions.

Proof. (optimum ⇒ KKT) The primal and dual feasibility conditions are satisfied by premise. From $g(\rho^\star, \lambda^\star) := \min_{x \in \mathbb{R}^d} L(x, \rho^\star, \lambda^\star)$ and strong duality $g(\rho^\star, \lambda^\star) = f(x^\star)$, it follows that $x^\star$ is an unrestricted minimum and must satisfy the stationarity condition $0 \in \partial_x L(x^\star, \rho^\star, \lambda^\star)$. From the implied equality $L(x^\star, \rho^\star, \lambda^\star) = f(x^\star)$ and feasibility, we must have the complementary slackness condition: $\rho_i^\star h_i(x^\star) = 0$ for all i. (KKT ⇒ optimum) This is Lemma 6.4. □

Combining Lemma 6.5 and Fact 6.1, we obtain Lemma 3.5, which we restate below as a corollary.

Corollary 6.6. Suppose that (1) $f, h_1 \ldots h_m, l_1 \ldots l_r$ are all affine, or (2) f and $h_1 \ldots h_m$ are convex; $l_1 \ldots l_r$ are affine; and there exists a strictly feasible point. Then $x^\star \in \mathcal{P}$ is a global minimum of f and $\rho^\star \geq 0_m$, $\lambda^\star \in \mathbb{R}^r$ are a global maximum of g iff they satisfies the KKT conditions.

6.2 SVM Dual

Let $X \in \mathbb{R}^{N \times d}$ and $y \in \{\pm 1\}^N$ where the i-th row $x_i \in \mathbb{R}^d$ of X represents a d-dimensional input vector and y_i is the corresponding binary label. Given some $C > 0$, the primal SVM problem is

$$w^\star, \xi^\star = \operatorname*{arg\,min}_{w \in \mathbb{R}^d,\ \xi \in \mathbb{R}^N:\ y \odot Xw \geq 1_N - \xi,\ \xi \geq 0_N} \frac{1}{2} \left|\left|w\right|\right|_2^2 + C \langle 1_N, \xi \rangle$$

Thanks to the slack variables, the primal problem is strictly feasible and strong duality holds. The Lagrangian is

$$L(w, \xi, \rho, \mu) = \frac{1}{2} \left|\left|w\right|\right|_2^2 + C \langle 1_N, \xi \rangle + \langle \rho, 1_N - \xi - y \odot Xw \rangle - \langle \mu, \xi \rangle$$

Since L is convex in $w \in \mathbb{R}^d$ and linear in $\xi \in \mathbb{R}^N$, and

$$\begin{aligned} \nabla_w L(w, \xi, \rho, \mu) = 0_d \qquad &\Leftrightarrow \qquad w = X^\top (\rho \odot y) \\ \nabla_\xi L(w, \xi, \rho, \mu) = 0_N \qquad &\Leftrightarrow \qquad \mu = C 1_N - \rho \end{aligned}$$

the Lagrangian dual function is

$$g(\rho, \mu) = \langle \rho, 1_N \rangle - \frac{1}{2} (\rho \odot y)^\top X X^\top (\rho \odot y)$$

While g is just a function of ρ, the constraint $\mu = C1_N - \rho \geq 0$ must be enforced. This creates the box constraint $0_N \leq \rho \leq C1_N$ as part of dual feasibility, and the dual problem is

$$\rho^\star = \operatorname*{arg\,max}_{0_N \leq \rho \leq C1_N} \ \langle \rho, 1_N \rangle - \frac{1}{2}(\rho \odot y)^\top XX^\top (\rho \odot y)$$

where $XX^\top \in \mathbb{R}^{N \times N}$ is the kernel matrix that enables kernelized SVMs. By complementary slackness,

$$\rho_i^\star (1 - \xi_i^\star - y_i \langle w^\star, x_i \rangle) = 0 \qquad \Leftrightarrow \qquad \rho_i^\star = 0 \quad \vee \quad y_i \langle w^\star, x_i \rangle = 1 - \xi_i^\star$$

Thus $\rho_i^\star > 0$ implies $y_i \langle w^\star, x_i \rangle \leq 1$: in this case x_i is called a support vector since $w^\star = \sum_{i=1:\ \rho_i^\star > 0}^N \rho_i^\star y_i x_i$. Furthermore, suppose $\rho_i^\star < C$. Note again by complementary slackness,

$$\mu_i^\star \xi_i^\star = 0 \qquad \Leftrightarrow \qquad \mu_i^\star = 0 \quad \vee \quad \xi_i^\star = 0$$

Thus $\mu_i^\star = C - \rho_i^\star > 0$ implies $\xi_i^\star = 0$. Combining these observations, we see that if $0 < \rho_i^\star < C$, then $y_i \langle w^\star, x_i \rangle = 1$.

6.3 Linear Programs

If the objective f and the constraints $h_1 \ldots h_m, l_1 \ldots l_r$ are all affine, we may standardize the problem as follows:

1. Write each variable as $x_i = x_i^+ - x_i^-$ where $x_i^+, x_i^- \geq 0$.

2. Write each equality constraint $a^\top x + \beta = 0$ as $a^\top x + \beta \geq 0$ and $-a^\top x - \beta \geq 0$.

3. Write each inequality constraint $a^\top x + \beta \leq 0$ as $a^\top x + \beta + s = 0$ where $s \geq 0$.

As a result, we can always write a linear program in the so-called canonical form:

$$f^* := \min_{x \geq 0_d:\ Ax \geq b} \ c^\top x \qquad c \in \mathbb{R}^d,\ A \in \mathbb{R}^{m \times d},\ b \in \mathbb{R}^m \tag{37}$$

The Lagrangian and the Lagrangian dual function are

$$\begin{aligned} L(x, \rho, \mu) &= c^\top x + \rho^\top (b - Ax) - \mu^\top x && \forall x \in \mathbb{R}^d,\ \rho \geq 0_m,\ \mu \geq 0_d \\ g(\rho, \mu) &= \min_{x \in \mathbb{R}^d} \ c^\top x + \rho^\top (b - Ax) - \mu^\top x && \forall \rho \geq 0_m,\ \mu \geq 0_d \end{aligned}$$

Given any $\rho \geq 0_m$ and $\mu \geq 0_d$, a minimizer $x \in \mathbb{R}^d$ of the Lagrangian must satisfy

$$\nabla_x L(x, \rho, \mu) = c - A^\top \rho - \mu = 0_d$$

Thus by fixing $c = A^\top \rho + \mu$ in the dual function, we have the dual problem

$$g^* = \max_{\rho \geq 0_m,\ \mu \geq 0_d:\ c = A^\top \rho + \mu} \ b^\top \rho = \max_{\rho \geq 0_m:\ A^\top \rho \leq c} \ b^\top \rho \tag{38}$$

where the last equality follows by treating $\mu \geq 0_d$ as a slack variable. Strong duality always holds in linear programs, so (37) and (38) are equivalent. This has a natural interpretation.

- In the primal (37), we buy d raw ingredients $x_1 \ldots x_d \geq 0$ to minimize the total cost $c^\top x \in \mathbb{R}$, while meeting minimum production bars $Ax \geq b$ in m products.

- In the dual (38), we assign prices $\rho_1 \ldots \rho_m \geq 0$ to m products to maximize the total profit $b^\top \rho$, while staying within our budget $A^\top \rho \leq c$ when purchasing d raw ingredients.

A linear program is a special case of semidefinite program (Section 7.1).

7 Matrix Extensions

Conic optimization generalizes convex analysis to *structured* inequalities (Appendix D). In this section, we focus on the special case of the Loewner order on symmetric matrices. We denote

$$\boldsymbol{S}^d := \left\{X \in \mathbb{R}^{d\times d} : X = X^\top\right\} \qquad \text{(vector space of symmetric matrices)}$$
$$\boldsymbol{S}^d_+ := \left\{X \in \boldsymbol{S}^d : X \succeq 0\right\} \qquad \text{(convex cone of PSD matrices)}$$

To impose structured inequality constraints, we assume $h_1 \dots h_m : \boldsymbol{S}^d \to \boldsymbol{S}^d$ defining some transformation of X inside $\boldsymbol{S}^d$ and assert $h_i(X) \preceq 0$. We assume r equality constraints $l_j(X) = 0$ defined by $l_1 \dots l_r : \boldsymbol{S}^d \to \mathbb{R}$. The hypothesis space is

$$\mathcal{P} = \left\{X \in \boldsymbol{S}^d : \ h(X) \preceq 0_m, \ l(X) = 0_r\right\} \tag{39}$$

where $\preceq$ is applied to each of the m matrices. Note that (39) is lacking scalar inequalities, but they add no modeling power and are thus often omitted from the standard form. Specifically, if $t : \boldsymbol{S}^d \to \mathbb{R}$ enforces a scalar inequality, we may write

$$t(X) \leq 0 \qquad \Leftrightarrow \qquad t(X) + s = 0, \ s \geq 0$$

where s can be absorbed into the system.[9] The primal problem is

$$f^\star := \min_{X\in\mathcal{P}} f(X) = \min_{X\in\mathcal{P}} \max_{\rho\in(\boldsymbol{S}^d_+)^m, \ \lambda\in\mathbb{R}^r} L(X,\rho,\lambda) \tag{40}$$

where the min-max formulation uses the Lagrangian

$$L(X,\rho,\lambda) = f(X) + \sum_{i=1}^{m} \langle \rho_i, h_i(X)\rangle + \sum_{j=1}^{r} \lambda_j l_j(X) \tag{41}$$

where $\langle A, B\rangle := \operatorname{tr}\left(A^\top B\right)$ is the matrix inner product. Prominently, each $\rho_i \in \boldsymbol{S}^d_+$ is itself a PSD matrix. To see why we may restrict the dual space to PSD matrices, note that

- If $B \preceq 0$, then $\langle A, B\rangle \leq 0$ for all $A \succeq 0$.[10]
- If $B \succ 0$, then $\langle A, B\rangle > 0$ for some $A \succ 0$.[11]

Thus if any inequality constraint is violated, the enemy can send the loss to infinity. The dual function and the dual problem are

$$g(\rho,\lambda) := \min_{X\in\boldsymbol{S}^d} L(X,\rho,\lambda) \tag{42}$$
$$g^\star := \max_{\rho\in(\boldsymbol{S}^d_+)^m, \ \lambda\in\mathbb{R}^r} g(\rho,\lambda) = \max_{\rho\in(\boldsymbol{S}^d_+)^m, \ \lambda\in\mathbb{R}^r} \min_{X\in\boldsymbol{S}^d} L(X,\rho,\lambda) \leq f^\star \tag{43}$$

where weak duality $g^\star \leq f^\star$ always holds. As before (Fact 6.1), strong duality holds under convexity + Slater (the same supporting-hyperplane argument goes through). Recall a matrix-valued function $h(X) \in \boldsymbol{S}^d$ is convex iff

$$h(\alpha X + (1-\alpha)Y) \preceq \alpha h(X) + (1-\alpha)h(Y)$$

for all $X, Y \in \boldsymbol{S}^d$ and $\alpha \in [0,1]$. However, linearity alone is no longer sufficient for strong duality (Section 7.1), so we only state the following.

[9] Alternatively, we may choose to keep it separate as usual (i.e., introduce a scalar Lagrangian multipler $\mu \geq 0$ for that inequality and add $\mu t(X)$ to the Lagrangian).

[10] Given $A, C \succeq 0$, we have $\langle A, C\rangle = \operatorname{tr}\left(A^{1/2}CA^{1/2}\right) = \left|\left|A^{1/2}C^{1/2}\right|\right|_F^2 \geq 0$. Set $C = -B$.

[11] E.g., take $A = vv^\top$ where $\lambda = v^\top Bv > 0$ is a positive eigenvalue of B.

Fact 7.1. Strong duality holds in (43) if

- f and $h_1 \dots h_m$ are convex; $l_1 \dots l_r$ are affine, and
- (Slater for generalized inequalities) There exists a strictly primal feasible point, namely $\bar{X} \in \boldsymbol{S}^d$ such that $h(\bar{X}) \prec 0_m$ and $l(\bar{X}) = 0_r$.

The new KKT conditions mirror the usual ones (Definition 3.1):

Definition 7.1. We say $X \in \mathbb{R}^{d\times d}$ satisfies the KKT conditions if we can find $\rho \in (\mathbb{R}^{d\times d})^m$ and $\lambda \in \mathbb{R}^r$ such that

1. **Primal feasibility**: $X \in \mathcal{P}$ (39)
2. **Dual feasibility**: $\rho_i \in \boldsymbol{S}^d_+$
3. **Complementary slackness**: $\langle \rho_i, h_i(X)\rangle = 0$ (equivalently, $\rho_i = 0_{d\times d}$ whenever $h_i(X) \prec 0$)
4. **Stationarity**: $\nabla_X L(X, \rho, \lambda) = 0_{d\times d}$

Analogously as in Lemma 6.5, if strong duality holds, convexity guarantees that an optimal solution is equivalent to a KKT point. Combining this with Fact 7.1, we summarize the result below:

Corollary 7.2. If (1) f and $h_1 \dots h_m$ are convex; $l_1 \dots l_r$ are affine, (2) there exists a strictly primal feasible point, then strong duality holds $g^\star = f^\star$ and attained by a KKT point $(X^\star, \rho^\star, \lambda^\star) \in \boldsymbol{S}^d \times (\boldsymbol{S}^d_+)^m \times \mathbb{R}^r$.

7.1 Semidefinite Programs (SDP)

A semidefinite program (SDP) generalizes a linear program (Section 6.3) from $\mathbb{R}^d$ to $\boldsymbol{S}^d$. Given $C, A_1 \dots A_r \in \boldsymbol{S}^d$ and $b \in \mathbb{R}^r$, it solves

$$f^\star := \min_{X \in \boldsymbol{S}^d:\ \langle A_j, X\rangle = b_j\ \forall j} \langle C, X\rangle \tag{44}$$

The Lagrangian is

$$L(X, \rho, \lambda) = \langle C, X\rangle - \langle \rho, X\rangle + \sum_{j=1}^r \lambda_j \left(b_j - \langle A_j, X\rangle\right) \tag{45}$$

Stationarity requires $\rho = C - \sum_{j=1}^r \lambda_j A_j$, implying the dual function

$$g(\rho, \lambda) = \begin{cases} b^\top \lambda & \text{if } \rho = C - \sum_{j=1}^r \lambda_j A_j \\ -\infty & \text{otherwise} \end{cases} \tag{46}$$

Folding in the dual feasibility $\rho \succeq 0$, we have the dual problem

$$g^\star = \max_{\lambda \in \mathbb{R}^r:\ \sum_{j=1}^r \lambda_j A_j \preceq C} b^\top \lambda$$

which is also an SDP.[12] Strong duality holds if (44) or (47) has a strictly feasible point (Slater's condition for the PSD cone). An SDP can be solved efficiently in practice, approximately in polynomial time (e.g., by the ellipsoid/interior-point methods).

[12] This uses an equivalent definition of SDP

$$\min_{\substack{x \in \mathbb{R}^K:\\ B_0 + x_1 B_1 + \cdots + x_K B_K \succeq 0}} c^\top x \tag{47}$$

where $B_0, B_1, \dots, B_K \in \boldsymbol{S}^d$ and $c \in \mathbb{R}^K$. It can be shown that (47) can be written as (44) by setting $X = B_0 + x_1 B_1 + \cdots + x_K B_K$ and taking $c_k = \langle C, B_k\rangle$.

7.2 AdaGrad Lemma

We use the following fact without proof.

Fact 7.3. If $P \succeq 0$, $f(X) = \mathrm{tr}\left(X^{-1}P\right)$ is convex over $X \succ 0$; strictly convex if $P \succ 0$.

Lemma 7.4 (Duchi *et al.* (2011), Lemma 15)**.** Pick $P \succeq 0$ and $c > 0$. Let

$$f^\star = \min_{X \in \boldsymbol{S}_+^d :\, \mathrm{tr}(X) \leq c} \mathrm{tr}\left(X^{-1}P\right) \tag{48}$$

If $P \succ 0$, the unique solution and its objective are as follows:

$$X^\star = \frac{c}{\mathrm{tr}\left(P^{1/2}\right)} P^{1/2} \qquad\qquad f^\star = \frac{1}{c}\mathrm{tr}\left(P^{1/2}\right)^2$$

If P is rank-deficient, (48) admits the same infimum $f^\star = c^{-1}\mathrm{tr}\left(P^{1/2}\right)$ but it is not attained by any $X \succ 0$.

Proof. We may assume $X \succ 0$ since the objective is undefined otherwise. Given Fact 7.3, clearly Corollary 7.2 holds (i.e., convex + Slater). The Lagrangian is

$$L(X, \rho, \mu) = \mathrm{tr}\left(X^{-1}P\right) - \langle \rho, X \rangle + \mu(\mathrm{tr}\,(X) - c)$$

Since $X \succ 0$, we immediately get $\rho = 0_{d\times d}$ by complementary slackness. Stationarity $-X^{-1}PX^{-1} - \mu I_d = 0_{d\times d}$ (and feasibility) then enforces that the solution and its objective must have the form

$$X = \frac{1}{\sqrt{\mu}} P^{1/2} \qquad\qquad f(X) = \sqrt{\mu}\,\mathrm{tr}\left(P^{1/2}\right)$$

which is unique since the PSD square root is unique. Since we want to minimize $f(X)$, we seek the smallest feasible $\mu \geq 0$. The constraint $\mathrm{tr}\,(X) \leq c$ implies $\sqrt{\mu} \geq (1/c)\mathrm{tr}\left(P^{1/2}\right)$. This yields the statement. For the rank-deficient case, we refer to the paper. □

References

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, **39**(3), 930–945.

Boyd, S. P. and Vandenberghe, L. (2014). *Convex Optimization*. Cambridge University Press.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, **12**(7).

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., *et al.* (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. (2023). Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, **36**, 50358–50376.

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*. Springer.

A Illustrations

A.1 Linearized Feasible Directions (Section 1.1)

Ball. The inequality constraint $h(x) = x_1^2 + x_2^2 - 1 \leq 0$ yields $\mathcal{P} = \{x \in \mathbb{R}^2 : x \text{ is inside the unit ball}\}$. At $x = (1, 0) \in \mathcal{P}$, the constraint is active (i.e., $h(1, 0) = 0$). Since $\nabla h(1, 0) = (2, 0)$, the linearized tangent cone is $T_{\text{linear}}(1, 0) = \{t \in \mathbb{R}^2 : t_1 \leq 0\}$ (i.e., the left half-plane). In contrast, the contraint is inactive at $x = (0, 0) \in \mathcal{P}$ (i.e., $h(0, 0) = -1 < 0$), thus $T(0, 0) = \mathbb{R}^2$.

Circle. The equality constraint $l(x) = x_1^2 + x_2^2 - 1 = 0$ yields $\mathcal{P} = \{x \in \mathbb{R}^2 : x \text{ lies on the unit circle}\}$. At $x = (1, 0) \in \mathcal{P}$, we have $\nabla l(1, 0) = (2, 0)$, thus $T_{\text{linear}}(1, 0) = \{t \in \mathbb{R}^2 : t_1 = 0\}$ (i.e., all vertical vectors).

Line. The equality constraint $l(x) = x_1 + x_2 = 0$ yields $\mathcal{P} = \{x \in \mathbb{R}^2 : x \text{ lies on the main diagonal line}\}$. At any $x \in \mathcal{P}$, we have $\nabla l(x) = (1, 1)$, thus $T_{\text{linear}}(x) = \{t \in \mathbb{R}^2 : t_1 + t_2 = 0\} = \mathcal{P}$.

A.2 KKT Conditions: Necessity (Section 3.1)

The loss to minimize is the linear plane $f(x) = x_1 + x_2$. The unit circle can be specified with the single equality constraint $l(x) = x^\top x - 1 = 0$. The Lagrangian is

$$L(x, \lambda) = 1_2^\top x + \lambda(x^\top x - 1)$$

For $x \in \mathbb{R}^2$ to be a KKT point, it must be feasible (i.e., $x^\top x = 1$) and have some $\lambda \in \mathbb{R}$ such that

$$\nabla_x L(x, \lambda) = 1_2 + 2\lambda x = 0_2$$

We see that $\lambda = 0$ fails, thus we may assume that $\lambda \neq 0$ and rewrite the stationarity condition as $x = (-\frac{1}{2\lambda}, -\frac{1}{2\lambda})$. Then since x must be feasible, we must have $x^\top x = \frac{1}{2\lambda^2} = 1$ or $\lambda = \pm\frac{\sqrt{2}}{2}$. Thus there are two KKT points:

$$x^{(1)} = \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) \qquad\qquad x^{(2)} = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$$

Now consider the unit ball, i.e., $x \in \mathbb{R}^2$ satisfying $h(x) = x^\top x - 1 \leq 0$. The Lagrangian is

$$L(x, \rho) = 1_2^\top x + \rho(x^\top x - 1)$$

For $x \in \mathbb{R}^2$ to be a KKT point, it must be feasible (i.e., $x^\top x \leq 1$) and have some $\rho \geq 0$ such that

$$\nabla_x L(x, \lambda) = 1_2 + 2\rho x = 0_2$$

Again, we rule out $\rho = 0$ and assume that $\rho > 0$. Then by complementary slackness, the constraint must be active. Rewriting the stationarity condition as $x = (-\frac{1}{2\rho}, -\frac{1}{2\rho})$ and solving for $\rho \in \mathbb{R}$ such that $x^\top x = 1$, we conclude that $\rho = \pm\frac{\sqrt{2}}{2}$. By dual feasibility, we conclude $\rho = \frac{\sqrt{2}}{2}$. Thus there is one KKT point:

$$x^{(1)} = \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right)$$

B Approximation Error of a Two-Layer Regressor

We consider regression $f : \mathbb{R}^d \to \mathbb{R}$ with squared loss $l(f(x), y) = (f(x) - y)^2$. Let $f^\star(x) = \mathbf{E}[Y|X = x]$ denote the Bayes-optimal regressor. Let

$$\mathcal{H}_N = \left\{ h_N(x) = \sum_{i=1}^{N} u_i \sigma(a_i^\top x + b_i) : \; a_1 \dots a_N \in \mathbb{R}^d, \; b, u \in \mathbb{R}^N \right\} \tag{49}$$

denote the hypothesis class of two-layer networks with N hidden units and activation σ (total $N(d+2)$ parameters). For a random unbiased predictor h (i.e., random weights), its approximation error can be written as

$$\mathbf{E}[L(h)] - L(f^\star) = \mathbf{E}[(h(X) - f^\star(X))^2] = \mathrm{Var}\,(h) \tag{50}$$

by the property of squared loss. Under mild boundedness assumptions, we can write $f^\star$ as an integral of single-neuron network ("Barron form"):

$$f^\star(x) = \int_{\mathbb{R}^d \times \mathbb{R}} s(a,b)\sigma(a^\top x + b)d(a,b) \tag{51}$$

with a finite "Barron norm" $B = \int_{\mathbb{R}^d \times \mathbb{R}} |s(a,b)|\, d(a,b) > 0$. (49) is viewed as a discrete version of (51). We craft a random hypothesis $h_N \in \mathcal{H}_N$ as follows:

$$\begin{aligned} p(a,b) &= \frac{|s(a,b)|}{B} & u_i &= \frac{B}{N}\mathbf{sign}(s(a_i,b_i)) \\ (a_1,b_1)\ldots(a_N,b_N) &\sim p & h_N(x) &= \sum_{i=1}^{N} u_i \sigma(a_i^\top x + b_i) \end{aligned}$$

This construction gives

$$\begin{aligned} \mathbf{E}[h_N(x)] &= B\mathbf{E}[\mathbf{sign}(s(a,b))\sigma(a^\top x + b)] \\ &= B\left(\int_{\mathbb{R}^d \times \mathbb{R}} \frac{|s(a,b)|\,\mathbf{sign}(s(a,b))}{B} \times \sigma(a^\top x + b)d(a,b)\right) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}} s(a,b)\sigma(a^\top x + b)d(a,b) \\ &= f^\star(x) \end{aligned}$$

Assuming $\mathrm{Var}\,\big(\sigma(a^\top x + b)\big) \leq C_x$ for $(a,b) \sim p$,

$$\mathrm{Var}\,(h_N(x)) \leq \frac{B^2 C_x}{N} \qquad \Rightarrow \qquad \mathrm{Var}\,(h_N) \leq \frac{C}{N} \qquad \big(C = B^2\mathbf{E}[C_X]\big)$$

Thus $h_N \to f^\star$ is a Monte-Carlo estimator of the Bayes-optimal regressor, getting more accurate with a wider width. Plugging it in (50) as an inferior to the best hypothesis $h_N^\star = \arg\min_{h_N \in \mathcal{H}_N} L(h_N)$, we have

$$L(h^\star) - L(f^\star) \leq \mathbf{E}[L(h_N)] - L(f^\star) \leq \frac{C}{N} \leq \frac{C}{\sqrt{N}} \tag{52}$$

Note that we actually have a stronger upper bound $O(N^{-1})$. But the weaker inverse square-root bound $O(N^{-1/2})$ (which is valid) is often invoked instead for convenience. The latter also appears when we analyze the nonasymptotic risk, e.g., by Chebyshev, defining the random excess $Z = L(h_N) - L(f^\star)$,

$$\Pr\left(Z > \frac{C}{\sqrt{N}}\right) < \frac{E[Z]}{C/\sqrt{N}} \leq \frac{C/N}{C/\sqrt{N}} = \frac{1}{\sqrt{N}}$$

C Chinchilla Approaches

The pioneering work of Chinchilla explores three approaches to modeling the relationship between the loss L, model size N, data size D, and compute budget C. A quick summary is

1. Create an envelop of loss curves across model/data sizes and predict the optimal $N \propto C^a$ and $D \propto C^b$.

2. Create isoFLOP curves across compute budgets (only using the final losses), use their minima to predict the optimal $N \propto C^a$ and $D \propto C^b$.

3. Directly model $L = O(N^{-\alpha} + D^{-\beta})$, which is fit on the training runs from Approach 1 and 2. This yields the optimal $N \propto C^a$ and $D \propto C^b$ subject to $C = \mathrm{FLOP}(N, D)$ (Section 5.1).

C.1 Approach 1

We first obtain 28 FLOP-to-loss curves $\phi_1 \ldots \phi_{28} : \mathbb{R} \to \mathbb{R}$ by running the following 28 workloads:

- For 7 model sizes $N \in \{0.075\text{B}, 0.25\text{B}, 0.5\text{B}, 1\text{B}, 2.5\text{B}, 5\text{B}, 10\text{B}\}$:
  - For 4 data sizes $D \in \left\{D_{\min}(N),\ 16^{1/3} D_{\min}(N),\ 16^{2/3} D_{\min}(N),\ 16 D_{\min}(N)\right\}$:
    - ∗ Train a N-parameter model on D tokens (cosine decay to 0.1 of the peak LR).
    - ∗ Smooth the resulting FLOP-to-loss curve to obtain $\phi : [0, C] \to \mathbb{R}$ where $C = \text{FLOP}(N, D)$.

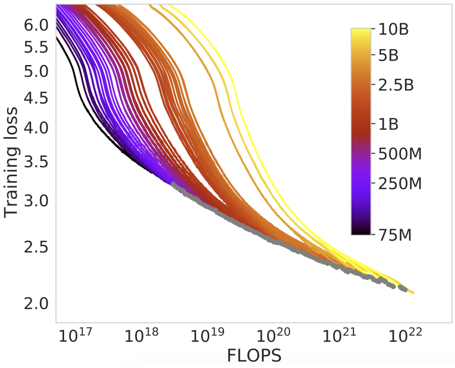
The compute-optimal envelope is then

$$\psi(C) = \min_{i=1}^{28} \phi_i(C)$$

We specify a data schedule that satisfies $C_{\max}(N) = \text{FLOP}(N, D_{\max}(N)) \geq C_{\min}(N') = \text{FLOP}(N', D_{\min}(N'))$ where $N' > N$ is the next model size. Since every loss curve starts from the same (untrained) loss at $C = 0$, this ensures that the envelope never has a "vertical gap". (It is still possible to have a change of slope at points where arg min changes.) Setting $D_{\min}(75\text{M}) = 7\text{B}$ as the smallest data size, and using a generous overlap $C_{\max}(N) = 6C_{\min}(N')$ with the heuristic $C = 6ND$, we can replicate the data points in the paper's plot (Figure 2, note the log-log scale):

N	D_1	D_2	D_3	D_4
0.075	7	18	44	112
0.25	6	14	36	90
0.5	7	19	47	119
1	10	25	63	159
2.5	11	27	67	170
5	14	36	90	227
10	19	48	120	302

(in billions)



The workloads can be executed in a reasonably quick period of time. Concretely, the largest workload would take roughly 2–3 days on a v5p-512 TPU cluster (assuming 170 TFLOP/s per core). Let $N(C), D(C)$ denote the optimal model and data size for compute C. We prepare 1500 labeled data points $(C_i, N(C_i), D(C_i))$ where C_i is log-spaced between 6×10^{18} and 2×10^{22} (corresponding to the gray dots in the plot), and fit

$$A^\star, a^\star = \operatorname*{arg\,min}_{A, a \in \mathbb{R}} \sum_{i=1}^{1500} (\log N(C_i) - \log A - a \log C_i)^2 \qquad B^\star, b^\star = \operatorname*{arg\,min}_{B, b \in \mathbb{R}} \sum_{i=1}^{1500} (\log D(C_i) - \log B - b \log C_i)^2$$

Then for any new compute C, we predict the optimal model and data sizes as $N^\star = A^\star C^{a^\star}$ and $D^\star = B^\star C^{b^\star}$. The paper estimates $a^\star = b^\star = 0.5$, suggesting that both the model and data sizes should grow at equal rate in compute, proportionally to $\sqrt{C}$. While the paper does not reveal the slopes $A^\star, B^\star$, it reveals the prediction for their compute budget $C = 5.76 \times 10^{23}$ as $N^\star = 67\text{B}$ and $D^\star = 1.5\text{T}$ (22.4 tokens per parameter), establishing the famous Chinchilla-optimal rule of thumb $D^\star \approx 20N^\star$.[13]

C.2 Approach 2

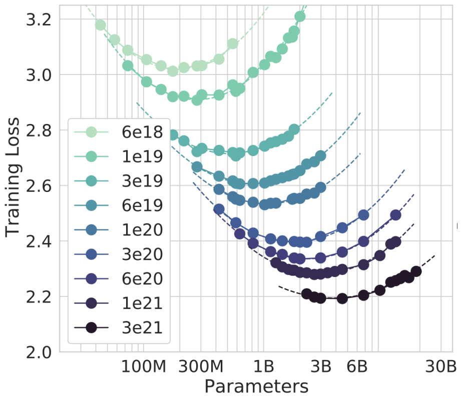
We first obtain losses(C) per compute budget C by running the following ≈99-135 workloads:

- For 9 FLOP budgets $C \in \{6, 10, 30, 60, 100, 300, 600, 1000, 3000\} \times 10^{18}$:
  - losses(C) = [ ]
  - For ≈11–15 model sizes $N \in \{0.07\text{B}, 0.22\text{B}, 0.7\text{B}, 1.5\text{B}, 2.2\text{B}, 3.5\text{B}, 5\text{B}, 7\text{B}, 10\text{B}, 12\text{B}, 16\text{B}\}$ (hypothetical):

[13] From the prediction, we can actually infer $A^\star \approx 0.088$ and $B^\star \approx 1.98$.

- Train a N-parameter model on $D = C/(6N)$ tokens (cosine decay to 0.1 of the peak LR).
- Take the final loss L_{final} (after smoothing).
- losses(C).append(L_{final})

The FLOP range is similar to Approach 1 (10^{18}–10^{22}). The model sizes are hypothetical as the paper does not reveal their values (other than they go up to 16B). We assume that they range from 70M to 16B in half-decade ($\sqrt{10}$, or $\frac{1}{2}$-step in log space) to ensure even spacing while having enough samples, and that they are adjusted as needed to ensure practicality (e.g., skip if $D = C/(6N)$ is too small/large) and clear visibility of a compute-optimal size (e.g., add a few fine-grained points around the predicted optimal $N^\star$ from Approach 1). These experiments yield isoFLOP curves as shown to the right (from the paper).



Since each isoFLOP curve looks like a parabola in log space, we fit $L \approx \beta_C(\log N - \alpha_C)^2 + \gamma_C$ for each C and use the minimum $N^\star = 10^{\alpha_C}$ and the corresponding data size $D^\star = C/(6N^\star)$ to create 9 labeled data points $(C, N^\star, D^\star)$. On these data points, we again fit power laws $N^\star \propto C^a$ and $D^\star \propto C^b$. The paper estimates $a^\star = 0.49$ and $b^\star = 0.51$. The paper again reveals the prediction for their compute budget: $N^\star = 63$B and $D^\star = 1.4$T (22.2 tokens per parameter).

C.3 Approach 3

Let $l : \Delta^{V-1} \times \mathcal{V} \to \mathbb{R}$ denote the next-word cross-entropy loss. Let **pop** denote a population distribution over context-word pairs $(X, Y) \in \mathcal{V}^T \times \mathcal{V}$ defining the true risk of a language model $f : \mathcal{V}^T \to \Delta^{V-1}$ as $L(f) = \mathbf{E}[l(f(X), Y)] \in \mathbb{R}$. Let $\mathcal{H}_N$ denote the hypothesis class of N-parameter transformers following a fixed blueprint. Let $(x_1, y_1) \dots (x_D, y_D) \sim$ **pop** denote D iid samples defining the empirical risk $\widehat{L}(f) = (1/D)\sum_{i=1}^{D} l(f(x_i), y_i)$. We have the usual players

$$f^\star = \arg\min_{f} \; L(f) \qquad \text{(Bayes optimal)}$$

$$h^\star = \arg\min_{h \in \mathcal{H}_N} \; L(h) \qquad \text{(best transformer)}$$

$$\hat{h} = \arg\min_{h \in \mathcal{H}_N} \; \widehat{L}(h) \qquad \text{(best finite-sample transformer)}$$

plus the actual $\tilde{h} \in \mathcal{H}_N$ we train in practice which has optimization flaws so that $\widehat{L}(\tilde{h}) > \widehat{L}(\hat{h})$. We have the error decomposition

$$L(\tilde{h}) = \underbrace{L(f^\star)}_{\text{irreducible error}} + \underbrace{L(h^\star) - L(f^\star)}_{\text{approximation error}} + \underbrace{L(\hat{h}) - L(h^\star)}_{\text{estimation error}} + \underbrace{L(\tilde{h}) - L(\hat{h})}_{\text{optimization error}}$$

Research on the universality of neural networks gives some guidance on the approximation error. For instance, optimal width-N two-layer regressors reduce the approximation error at a dimension-free rate of $O(N^{-1/2})$ (Barron, 1993), see Appendix B for a proof sketch. The estimation error is the expected regret (of a single hypothesis rather than an online learning algorithm), which has the well-known lower bound of $O(D^{-1/2})$. Thus Chinchilla defines

$$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} \tag{53}$$

to predict the risk of an N-parameter transformer trained on D tokens. The parameters $\alpha, \beta, A, B, E \in \mathbb{R}$ are regressed on the training runs from Approach 1 and 2 (total $n \approx 130$–160 labeled points). To combat outliers, Chinchilla uses the Huber loss (i.e., squared loss for residuals $\leq \delta$, appropriately scaled absolute loss for larger residuals); tiny $\delta = 0.001$ is found to be necessary to handle sensitive residuals. Since N, D are large values, (53) is computed in log space for numerical stability using the log-sum-exp trick: define $e = \log E$, $a = \log A$, and $b = \log B$ and write

$$\begin{aligned}\log L(N, D) &= \log\left(\exp(e) + \exp(a - \alpha \log N) + \exp(b - \beta \log D)\right)\\ &= \mu + \log\left(\exp(e - \mu) + \exp(a - \alpha \log N - \mu) + \exp(b - \beta \log D - \mu)\right)\\ &= \mathrm{LSE}(e, a - \alpha \log N, b - \beta \log D)\end{aligned}$$

where $\mu = \max(e, a - \alpha \log N, b - \beta \log D)$. Thus the optimization problem is

$$\alpha^\star, \beta^\star, a^\star, b^\star, e^\star = \underset{\alpha,\beta,a,b,e\in\mathbb{R}}{\arg\min} \sum_{i=1}^{n} \mathrm{Huber}_{\delta=0.001}(\log L_i - \mathrm{LSE}(e, a - \alpha \log N_i, b - \beta \log D_i)) \tag{54}$$

where we can recover $E^\star = \exp(e^\star)$, $A^\star = \exp(a^\star)$, and $B^\star = \exp(b^\star)$. Since this is a 5-dimensional problem on a couple hundred points, we can use computationally costly algorithms (the paper uses LBFGS, but even that is not necessary, e.g., we can use exact Newton). The problem is nonconvex, so Chinchilla sweeps a grid of values for initialization that is large enough to strictly contain the optimal initialization inside the grid. Chinchilla reports $E^\star = 1.69$, $A^\star = 406.4$, $B^\star = 410.7$, $\alpha^\star = 0.34$, and $\beta^\star = 0.28$. The finding $\beta^\star < 0.5$ is consistent with the estimation error lower bound (i.e., the risk decays slower than what is possible). The finding $\alpha^\star < 0.5$ is inconsistent with the approximation upper bound, but the theorem applies to a different architecture and loss.

C.3.1 Data-constrained extension

Chinchilla assumes that the D tokens carry the same amount of information (i.e., they are iid samples). Muennighoff *et al.* (2023) propose to model data epoching by assuming that

$$D = U + (1-\delta)U + (1-\delta)^2 U + \cdots + (1-\delta)^{R_D} U \tag{55}$$

where U is the number of "unique" tokens, R_D the number of epochs, and $\delta \in (0, 1)$ a discount factor.[14] We express (55) as $D \approx U + \alpha U$ for interpretability. Taking the geometric sum and defining $R_D^\infty := \frac{1-\delta}{\delta}$, we have

$$D = U + R_D^\infty \left(1 - (1-\delta)^{R_D}\right) U \tag{56}$$

where $D = U + R_D^\infty U$ in the limit $R_D \to \infty$ (i.e., the most we can "squeeze out" of U tokens by epoching is $R_D^\infty U$ additional tokens). The authors further simplify the form as a function of $\frac{R_D}{R_D^\infty}$ by assuming that $\delta \approx 0$. In this case, $\delta = \frac{1}{R_D^\infty + 1} \approx \frac{1}{R_D^\infty}$ and $1 - \delta \approx e^{-\delta}$, which imply

$$D \approx U + R_D^\infty \left(1 - e^{-\frac{R_D}{R_D^\infty}}\right) U \tag{57}$$

Similarly, Chinchilla assumes that the N parameters are equally valuable, but in practice they have decreasing marginal utility. For instance, with $U = 10$ tokens, increasing the model size from from 1 to 10 parameters reduces the loss much faster than from 101 to 110. The authors use a symmetric parameterization to model excess parameters:

$$N \approx N(U) + R_N^\infty \left(1 - e^{-\frac{R_N}{R_N^\infty}}\right) N(U) \tag{58}$$

where $N(U) = G(UG)^{\alpha/\beta}$ is the Chinchilla-optimal model size for U unique tokens (30) and R_N is the "model epochs", yielding $N = N(U) + R_N^\infty N(U)$ in the limit $R_N \to \infty$. Plugging these in the Chinchilla loss (53), we have

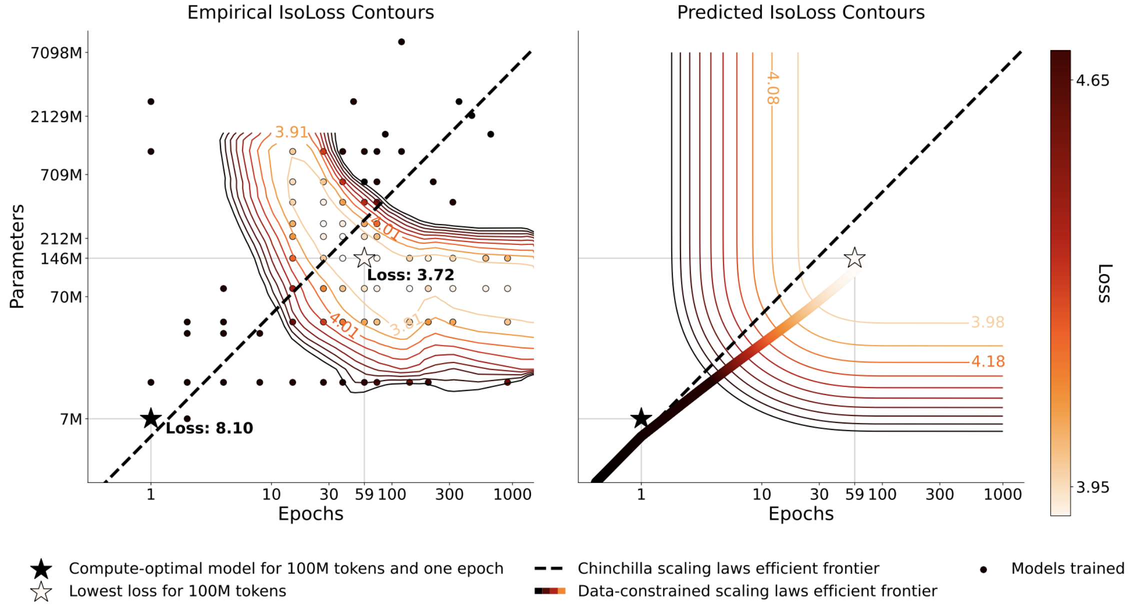
$$L(U, R_D, R_N) = E + \frac{A}{\left(N(U) + R_N^\infty \left(1 - e^{-\frac{R_N}{R_N^\infty}}\right) N(U)\right)^\alpha} + \frac{B}{\left(U + R_D^\infty \left(1 - e^{-\frac{R_D}{R_D^\infty}}\right) U\right)^\beta} \tag{59}$$

[14] Note that even this is a vast simplification of the real-world setting where we use a data mixture with different epoching specifications (e.g., 20 epochs on Wikipedia, but 0.3 epoch on Common Crawl). Since real pretraining does not neatly decompose into epoching over U tokens, it is not entirely clear how to use this extension for practical purposes. The paper addresses this point by empirically showing that inherent duplication in U does not change the optimal epoch. Note also that a major "bug" in this formulation is that epoching can only further reduce the loss, which fails to model overfitting. The paper addresses this point by decaying the exponent β instead of increasing D, but it results in a poor fit

which can be used to predict losses corresponding to unique data, data epochs, and model epochs. We fit (59) in two steps. First, $\alpha = \beta, a, b, e$ are learned by (54) on the U unique tokens, resulting in $N(U) = 0.051U$ (or $U = 19.6N(U)$). Then R_D^∞, R_N^∞ are learned over $n = 182$ samples of $(U_i, R_{D,i}, R_{N,i}, L_i)$ where the unique tokens U and the data/model epochs R_D, R_N are varied (1–500 epochs) covering model sizes (58) from 7M to 9B:

$$\begin{aligned} &\min_{R_D^\infty, R_N^\infty \in \mathbb{R}} \sum_{i=1}^{n} \text{Huber}_{\delta=0.001}\Bigg(\log L_i - \\ &\text{LSE}\left(e, a - \alpha \log\left(N(U_i) + R_N^\infty\left(1 - e^{-\frac{R_{N,i}}{R_N^\infty}}\right)N(U_i)\right), b - \beta\log\left(U_i + R_D^\infty\left(1 - e^{-\frac{R_{D,i}}{R_D^\infty}}\right)U_i\right)\right)\Bigg) \qquad (60) \end{aligned}$$

Given the flawed formulation that does not allow for nonmonotonic loss reduction, the samples are noisy (e.g., double descent with data epoching). The authors nonetheless fit $R_D^\infty = 15.4$ and $R_N^\infty = 5.3$, suggesting that we squeeze out more from data epoching than model epoching. Here is a plot from the paper:



The data points correspond to training on $U = 100$M unique tokens with epoching ($N(U) = 5.1$M). The isoloss contours are interpolated from the data points on the left and predicted by (59) on the right; in the latter case, the convex shape is because epoching cannot increase the loss. Both (53) and (59) reasonably predict the location of optimal model-data without epoching (black star), but the latter becomes more accurate with epoching (white star) because it models diminishing returns. A limitation is that it does not give a closed-form solution for compute-optimal U, R_D, R_N, unlike Chinchilla which gives a closed-form solution for compute-optimal N, D (the above frontier is drawn by plotting), but one may consider numerically minimizing (59) subject to compute constraints (e.g., grid search). The paper finds that for fixed U, the loss predicted by (59) closely matches the Chinchilla loss up to $R_D = 4$ data epochs across different model sizes (i.e., up to 4 epochs, the repeated data is as good as new for the purpose of reducing the test loss).

D Conic Optimization

D.1 Convex Cones and Generalized Inequality

A **cone** $K \subseteq V$ is a subset of an inner product vector space V such that $x \in K$ implies $\alpha x \in K$ for all $\alpha \geq 0$. The **dual cone** of K is $K^* := \{y \in V : \langle x, y\rangle \geq 0 \; \forall x \in K\}$. A cone K is **convex** if every conic (i.e., nonnegative linear) combination is contained in K. A cone K is **proper** if convex, closed, $\text{int}(K) \neq \varnothing$, and "pointed" ($-x, x \in K$ implies $x = 0$). A proper cone K defines a **generalized inequality** by

$$x \preceq_K y \iff y - x \in K$$
$$x \prec_K y \iff y - x \in \text{int}(K)$$

It preserves most properties of ordinary inequality (e.g., transitive), but it is a partial ordering over V (e.g., not every pair is comparable). This makes the notion of minimum and maximum subtle (e.g., minimum $\neq$ minimal, no least upper bounds). Given a proper cone $K \subseteq V$, a function $f : \mathcal{X} \to V$ is called **K-convex** if $f(\alpha x + (1-\alpha)y) \preceq_K \alpha f(x) + (1-\alpha) f(y)$ for all $x, y \in \mathcal{X}$ and $\alpha \in [0,1]$.

Examples

- A subspace $S \subseteq \mathbb{R}^d$ is a cone. Its dual cone is the orthogonal complement $S^\perp$ (why?).
- The nonnegative orthant $\mathbb{R}^d_{\geq 0} \subset \mathbb{R}^d$ is a proper cone. Its dual cone is itself ("self-dual"). The associated inequality is componentwise inequality.
- The set of PSD matrices $\boldsymbol{S}^d_+ \subset \boldsymbol{S}^d$ is a proper cone. Its dual cone is itself (again, self-dual)—we leave the proof as a quick exercise. The associated generalized inequality is Loewner order. Section 7 focuses on this special case.

D.2 Generalized KKT Conditions

Assume an objective function $f : \mathcal{X} \to \mathbb{R}$ for some vector space $\mathcal{X}$, $h_i : \mathcal{X} \to V_i$ corresponding to a proper cone $K_i \subseteq V_i$ for $i = 1 \ldots m$, and $l_j : \mathcal{X} \to \mathbb{R}$ for $j = 1 \ldots r$. Consider

$$f^\star := \min_{\substack{x \in \mathcal{X}: \\ h_i(x) \preceq_{K_i} 0 \ \forall i = 1 \ldots m \\ l_j(x) = 0 \ \forall j = 1 \ldots r}} f(x) \tag{61}$$

The corresponding Lagrangian is

$$L(x, \rho = \{\rho_i\}_{i=1}^m, \lambda) = f(x) + \sum_{i=1}^m \langle \rho_i, h_i(x) \rangle + \sum_{j=1}^r \lambda_j l_j(x) \tag{62}$$

where $\rho_i \in V_i$.

Definition D.1. If $(x^\star, \rho^\star, \lambda^\star) \in \mathcal{X} \times (\bigtimes_{i=1}^m V_i) \times \mathbb{R}^r$ satisfy

1. **Primal feasibility**: $h_i(x^\star) \preceq_{K_i} 0$ for all i and $l(x^\star) = 0_r$
2. **Dual feasibility**: $\rho_i^\star \succeq_{K_i^*} 0$ for all i
3. **Stationarity**: $\nabla_x L(x^\star, \rho^\star, \lambda^\star) = 0$
4. **Complementary slackness**: $\langle \rho_i, h_i(x) \rangle = 0$ for all $i = 1 \ldots m$

then we say $(x^\star, \rho^\star, \lambda^\star)$ satisfy the KKT conditions.

D.3 Generalized Duality

For the dual function is $g(\rho, \lambda) := \min_{x \in \mathcal{X}} L(x, \rho, \lambda)$, the dual problem is

$$g^\star := \max_{\rho_i \in V_i : \rho_i \succeq_{K_i^*} 0, \ \lambda \in \mathbb{R}^r} g(\rho, \lambda) \tag{63}$$

where weak duality $g^\star \leq f^\star$ holds.

Lemma D.1. If

1. $f : \mathcal{X} \to \mathbb{R}$ is convex, $h_i : \mathcal{X} \to V_i$ is K_i-convex for $i = 1 \ldots m$; $l_1 \ldots l_r$ are affine, and
2. There exists a strictly primal feasible point, namely $x \in \mathcal{X}$ such that $h_i(x) \prec_{K_i} 0$ and $l(x) = 0_r$,

then strong duality holds $g^\star = f^\star$ and the dual optimum is attained.

E Proofs and Lemmas

Proof of Lemma 1.1. Pick any $t \in T(x)$.

1. Pick any $i \in I(x)$. The definition (2) implies that for all sufficiently small $\eta > 0$

$$h_i(x + \eta t) = h_i(x) + \eta \nabla h_i(x)^\top t + o(\eta) \leq 0 \tag{64}$$

   Since $h_i(x) = 0$, (64) is equivalent to $\nabla h_i(x)^\top t + \frac{o(\eta)}{\eta} \leq 0$. Thus we must have

$$\lim_{\eta \to 0^+} \nabla h_i(x)^\top t + \frac{o(\eta)}{\eta} = \nabla h_i(x)^\top t \leq 0$$

2. Pick any j. Similarly by definition, $t \in T(x)$ satisfies for all sufficiently small $\eta > 0$

$$l_j(x + \eta t) = l_j(x) + \eta \nabla l_j(x)^\top t + o(\eta) = 0 \tag{65}$$

   Since $l_j(x) = 0$, (65) is equivalent to $\nabla l_j(x)^\top t + \frac{o(\eta)}{\eta} = 0$. Thus we must have

$$\lim_{\eta \to 0^+} \nabla l_j(x)^\top t + \frac{o(\eta)}{\eta} = \nabla l_j(x)^\top t = 0$$

In conclusion, $t \in T_{\text{linear}}(x)$. □

Proof of Lemma 1.2. Pick any $t \in T_{\text{linear}}(x)$.

1. Pick any $i \notin I(x)$. Then $h_i(x + \eta t) = h_i(x) + O(\eta) \leq 0$ for all small enough $\eta > 0$ since $h_i(x) < 0$.
2. Pick any $i \in I(x)$. Then $h_i(x + \eta t) \leq h_i(x) + \eta \nabla h_i(x)^\top t \leq \eta \nabla h_i(x)^\top t \leq 0$ for all small enough $\eta > 0$. The first inequality is because h_i is locally concave, the second $i \in I(x)$, and the third $t \in T_{\text{linear}}(x)$.
3. Pick any j. Then $l_j(x + \eta t) = l_j(x) + \eta \nabla l_j(x)^\top t = 0$ for all small enough $\eta > 0$. The first inequality is because l_j is locally affine, and the second equality is because $x \in \mathcal{P}$ and $t \in T_{\text{linear}}(x)$.

Thus $t \in T(x)$. □

Proof of Lemma 1.3. Pick any $t \in T_{\text{linear}}(x)$. We want to show the existence of a limiting sequence

$$z(\eta) = x + \eta t + o(\eta) \tag{66}$$

where $z(0) = x$. Then since $\frac{z(\eta) - x}{\eta} = t + O(1)$,

$$\lim_{\eta \to 0^+} \frac{z(\eta) - x}{\eta} = t$$

This validates $t \in T_{\text{linear}}(x)$ as a genuine feasible direction, i.e., $t \in T(x)$. Let $m_x = |I(x)| \leq m$ denote the number of active inequality constraints at x (WLOG, we assume the first m_x are active). Let $c_x : \mathbb{R}^d \to \mathbb{R}^{m_x + r}$ evaluate these constraints and $A_x \in \mathbb{R}^{(m_x + r) \times d}$ denote the gradients. Specifically,

$$c_x(z) = \begin{bmatrix} h_1(z) \\ \vdots \\ h_{m_x}(z) \\ l_1(z) \\ \vdots \\ l_r(z) \end{bmatrix} \in \mathbb{R}^{m_x + r} \qquad A_x = \begin{bmatrix} \nabla h_1(x)^\top \\ \vdots \\ \nabla h_{m_x}(x)^\top \\ \nabla l_1(x)^\top \\ \vdots \\ \nabla l_r(x)^\top \end{bmatrix} \in \mathbb{R}^{(m_x + r) \times d} \tag{67}$$

where $c_x(x) = 0_{m_x + r}$ and $\nabla c_x(x) = A_x$ by definition. Now we define a helper function $R_x : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ by

$$R_x(z, \eta) = \begin{bmatrix} c_x(z) \\ V^\top (z - (x + \eta t)) \end{bmatrix} \in \mathbb{R}^d \tag{68}$$

where $z - (x + \eta t) \in \mathbb{R}^d$ is the correction needed for z to match $x + \eta t$ (see (66)) and $V \in \mathbb{R}^{d \times (d - m_x - r)}$ is an orthonormal basis of $\text{null}\,(A_x) \subset \mathbb{R}^d$. Note that

$$R_x(z, \eta) = 0_d$$

enforces *two* conditions. First, z is feasible (since $c_x(z) = 0_{m_x + r}$). Second, $z - (x + \eta t) \in \text{row}\,(A_x)$ (since $\text{null}\,(A_x) \perp \text{row}\,(A_x)$). Intuitively, the second condition allows for a "wiggle room" for the correction to achieve its goal. At any η, the Jacobian of R_x with respect to z is

$$\nabla_z R_x(z, \eta) = \begin{bmatrix} \nabla c_x(z) \\ V^\top \end{bmatrix} \in \mathbb{R}^{d \times d} \tag{69}$$

Evaluating the Jacobian at $z = x$ and $\eta = 0$, we have

$$\nabla_z R_x(x, 0) = \begin{bmatrix} A_x \\ V^\top \end{bmatrix} \in \mathbb{R}^{d \times d} \tag{70}$$

By LICQ at x, $A_x \in \mathbb{R}^{(m_x + r) \times d}$ is full-rank (see (67)). By definition, the columns of $V \in \mathbb{R}^{d \times (d - m_x - r)}$ are linearly independent of the rows of A_x. Therefore, (70) is an invertible $d \times d$ matrix. Now we invoke the implicit function theorem: since

- $R_x(z, \eta)$ is continuously differentiable.
- $R_x(x, 0) = 0_d$
- $\nabla_z R_x(x, 0) \in \mathbb{R}^{d \times d}$ is invertible.

there exists a smooth sequence $z(\eta) \in \mathbb{R}^d$ in η around 0 such that

1. $z(0) = x$
2. $R_x(z(\eta), \eta) = 0_d$ for all η sufficiently close to 0

The final step is to show that this sequence satisfies (66). Since $z : \mathbb{R} \to \mathbb{R}^d$ is elementwise smooth in $\eta \in \mathbb{R}$, let $z' : \mathbb{R} \to \mathbb{R}^d$ denote the Jacobian (i.e., $z_i'(\eta) = \frac{\partial z_i(\eta)}{\partial \eta}$).

Lemma. $z'(0) = t$
(66) follows from the lemma since

$$z(\eta) = z(0) + \eta z'(0) + o(\eta) = x + \eta t + o(\eta)$$

Proof of the lemma. Denote the Jacobian of $R_x(z(\eta), \eta) \in \mathbb{R}^d$ with respect to $\eta \in \mathbb{R}$ by

$$J(\eta) = \nabla_z R_x(z(\eta), \eta) z'(\eta) + \nabla_\eta R_x(z(\eta), \eta)$$

which uses the chain rule. We have $\nabla_z R_x(z(\eta), \eta) = (\nabla c_x(z(\eta)), V^\top) \in \mathbb{R}^{d \times d}$ from (69). For the second term, we see $\nabla_\eta R_x(z(\eta), \eta) = (0_{m_x + r}, -V^\top t) \in \mathbb{R}^d$ from (68). Thus at $\eta = 0$, since $z(0) = x$ by Condition 1,

$$J(0) = \nabla_z R_x(z(0), 0) z'(0) + \nabla_\eta R_x(z(0), 0) = \begin{bmatrix} A_x z'(0) \\ V^\top (z'(0) - t) \end{bmatrix}$$

At the same time, since $R_x(z(\eta), \eta) = 0_d$ for η sufficiently close to 0 by Condition 2, we must have $J(\eta) = 0_d$ (differentiate both sides with respect to η) identically for that neighborhood. This implies $J(0) = 0_d$. This condition states that (i) $z'(0) \in \text{null}\,(A_x)$, and (ii) $z'(0) - t \in \text{row}\,(A_x)$. The only way to satisfy this is $z'(0) = t$ since $\text{null}\,(A_x) \perp \text{row}\,(A_x)$. □

Proof of Fact 2.1 and 2.2. Fact 2.1 has to be true if we were to have $f(x + \eta t) = f(x)$ for arbitrarily small $\eta > 0$ in (8). To see Fact 2.2, by premise there is some $\epsilon > 0$ such that $f(x + \eta t) - f(x) = \sum_{i=1}^{\infty} \frac{\eta^i}{i!} \nabla^i f(x).\textbf{contract}(t) > 0$ for all $\eta \in (0, \epsilon)$. This implies that it cannot be $\nabla^i f(x).\textbf{contract}(t) = 0$ for all $i \in \mathbb{N}$. Let k be the index of the first nonzero term. Then $f(x + \eta t) - f(x) = \frac{\eta^k}{k!} \nabla^k f(x).\textbf{contract}(t) + R_{k+1}(x + \eta t)$ for all $\eta \in (0, \epsilon)$. Since this must stay positive and $R_{k+1}(x + \eta t) = o(\eta^k)$, we must have $\nabla^k f(x).\textbf{contract}(t) > 0$. □

Proof of Lemma 2.3. Pick any $t \in T(x)$. If f is locally constant in t, from Fact 2.1 we trivially have $f^{(K)}(x+\eta t) = f^{(K)}(x)$ for all $\eta \in \mathbb{R}$. Now assume f is locally increasing in t. Since $f(x+\eta t) = f^{(K)}(x+\eta t) + R_{K+1}(x+\eta t)$ and $R_{K+1}(x) = 0$,

$$f^{(K)}(x+\eta t) - f^{(K)}(x) = (f(x+\eta t) - f(x)) - R_{K+1}(x+\eta t) \tag{71}$$

By Fact 2.2, there is some $k \in \mathbb{N}$ such that $f(x+\eta t) - f(x) = \frac{\eta^k}{k!}\nabla^k f(x).\mathbf{contract}(t) + R_{k+1}(x+\eta t)$ for all small enough $\eta > 0$ where $\nabla^k f(x).\mathbf{contract}(t) > 0$. Plugging this in (71), we have

$$f^{(K)}(x+\eta t) - f^{(K)}(x) = \frac{\eta^k}{k!}\nabla^k f(x).\mathbf{contract}(t) > 0$$

for all small enough $\eta > 0$. □

Proof of Lemma 2.4. Rearranging (71), we have

$$f(x+\eta t) - f(x) = \left(f^{(K)}(x+\eta t) - f^{(K)}(x)\right) + R_{K+1}(x+\eta t)$$

Applying Fact 2.2 to $f^{(K)}$, there is some $k \leq K$ such that $f^{(K)}(x+\eta t) - f^{(K)}(x) = \frac{\eta^k}{k!}\nabla^k f(x).\mathbf{contract}(t) + R_{k+1}(x+\eta t)$ for all small enough $\eta > 0$ where $\nabla^k f(x).\mathbf{contract}(t) > 0$. Thus

$$f(x+\eta t) - f(x) = \frac{\eta^k}{k!}\nabla^k f(x).\mathbf{contract}(t) + R_{k+1}(x+\eta t) + R_{K+1}(x+\eta t)$$

for all small enough $\eta > 0$. Since the remainder terms are $o(\eta^k)$ and the non-remainder term is $\eta^k c$ for some $c > 0$, the RHS is eventually positive (equivalently $f(x+\eta t) > f(x)$) for all small enough $\eta > 0$. □

Proof of Corollary 2.5. Since $x \in \mathcal{P}$ is a strict local minimum of $f^{(K)}$ over $\mathcal{P}$, there exist $\epsilon_1, \delta > 0$ such that for all $\eta \in (0, \epsilon_1)$

$$f^{(K)}(x+\eta t) \geq f^{(K)}(x) + \delta \qquad \forall t \in T(x)$$

Since $\left|f(x+\eta t) - f^{(K)}(x+\eta t)\right| = o(\eta^K)$, we can find some $\epsilon_2 > 0$ small enough to ensure that for all $\eta \in (0, \epsilon_2)$

$$f^{(K)}(x+\eta t) - f(x+\eta t) < \frac{\delta}{2}$$

Let $\epsilon = \min(\epsilon_1, \epsilon_2)$. Then for all $\eta \in (0, \epsilon)$,

$$f(x+\eta t) > f^{(K)}(x+\eta t) - \frac{\delta}{2} \geq f^{(K)}(x) + \frac{\delta}{2} > f^{(K)}(x) = f(x) \qquad \forall t \in T(x)$$

Thus x is a strict local minimum of f over $\mathcal{P}$. □

Proof of Lemma 4.1. Since $C > 0$, the origin is strictly feasible, thus Slater's condition holds (Lemma 3.5). The Lagrangian is

$$L(x, \rho) = g^\top x + \rho(x^\top A x - C^2)$$

where $\rho \geq 0$. Stationarity gives us $g + 2\rho A x = 0$. Since $g \neq 0_d$, we cannot have $\rho = 0$, thus it must be that $\rho > 0$ and we can write

$$x = -\frac{1}{2\rho}A^{-1}g \tag{72}$$

Since $\rho > 0$, by complementary slackness x must be on the boundary. Solving for $\rho > 0$ in $x^\top A x = C^2$ gives us

$$\rho = \frac{||g||_{A^{-1}}}{2C}$$

Plugging it in (72), we have the statement.

(An alternative proof is to first argue that (23) is achieved at the boundary $||x||_A = C$ and convert the inequality constraint to equality without changing the problem. While it simplifies the problem to solving a system of $d+1$ equalities, the feasible set is now a (nonconvex) circle and we have to demonstrate the sufficiency of KKT (e.g., by LICQ) and collect the minimum from the two resulting KKT points (Figure 5 left).) □

Proof of Lemma 4.2. The objective is convex and Slater's condition holds. The Lagrangian is

$$L(x, \rho) = x^\top Hx + \rho(x^\top x - 1)$$

Stationarity gives us $(H + \rho I_d)x = 0_d$. If $\rho > 0$, it requires the boundary condition $||x||_2 = 1$, but $H + \rho I_d$ is invertible so that $x = 0_d$ (i.e., $\rho > 0$ is impossible). If $\rho = 0$, we have $Hx = 0_d$, thus any $x \in \text{null}\,(H)$ with $||x||_2 \leq 1$ satisfies the condition and is a global minimum. The achieved objective is $f^\star = 0$. □

Proof of Lemma 4.3. Since $f^\star \leq \lambda_d < 0$, the solution must lie on the boundary. Thus we can write (24) as

$$f^\star = \min_{x \in \mathbb{R}^d :\, ||x||_2 = 1} \quad x^\top Hx \tag{73}$$

The equality constraint $l(x) = x^\top x - 1$ has the gradient $\nabla l(x) = 2x$, which is nonzero on the whole feasible set $\mathcal{S}^{d-1}$ (unit sphere), thus LICQ holds and all local minima of (73) are KKT points. The Lagrangian is

$$L(x, \lambda) = x^\top Hx + \lambda(1 - x^\top x)$$

Stationarity gives us $Hx = \lambda x$. By primal feasibility $x \in \mathcal{S}^{d-1}$, we have that x is a KKT point iff x is a unit-length eigenvector of H with an eigenvalue $\lambda \in \mathbb{R}$. Thus v_d must be a global minimum in (73) (hence (24)), achieving $f^\star = \lambda_d < 0$. □

Proof of Lemma 4.4. Since LICQ holds, the feasible directions at v_i are given by the linearized tangent cone $T_{\text{linear}}(v_i) = \left\{u \in \mathbb{R}^d : u^\top v_i = 0\right\}$, which includes v_j for all $j \neq i$. Taking a step $\eta > 0$ from v_i in v_j takes us to

$$c(\eta) = \frac{v_i + \eta v_j}{||v_i + \eta v_j||_2} = \frac{v_i + \eta v_j}{\sqrt{1 + \eta^2}} \in \mathcal{S}^{d-1}$$

It is easy to verify that $c(0) = v_i$ and $c'(0) = v_j$. The loss at $c(\eta)$ is

$$f(c(\eta)) = c(\eta)^\top Hc(\eta) = \lambda_i + \eta^2(\lambda_j - \lambda_i) + O(\eta^4) \qquad \Rightarrow \qquad \lim_{\eta \to 0^+} \; f(c(\eta)) \begin{cases} > f(v_i) & \text{if } \lambda_j > \lambda_i \\ = f(v_i) & \text{if } \lambda_j = \lambda_i \\ < f(v_i) & \text{if } \lambda_j < \lambda_i \end{cases}$$

That is, an infinitesimal feasible step from v_i in v_j can increase, decrease, or untouch the objective depending on the gap $\lambda_j - \lambda_i$. □

Proof of Lemma 4.5. Since (26) is convex and satisfies Slater's condition, any minimizer x satisfies stationarity $0_d \in \nabla f(x) + \rho\partial\,||x||$ for some multiplier $\rho \geq 0$. Since the minimizer $x^\star$ is nonstationary and nonzero, its associated multiplier $\rho^\star > 0$ must be positive. On the other hand, a minimizer of (27) is any stationary point $x \in \mathbb{R}^d$ satisfying $0_d \in \nabla f(x) + D\partial\,||x||$. Setting $D = \rho^\star$ gives the statement. □

Proof of Lemma 4.6. We know $0_d \in \nabla f(x^\star) + D\partial\,||x^\star||$. Since $x^\star$ is nonzero, $\nabla f(x^\star) \neq 0_d$. Since (26) is convex and satisfies Slater's condition, x is a minimizer iff it satisfies $0_d \in \nabla f(x) + \rho\partial\,||x||$ for some multiplier $\rho \geq 0$ such that $\rho(||x|| - C) = 0$. Setting $\rho = D$ and $C = ||x^\star||$, we conclude that $x^\star$ is a minimizer. □