CS395T: Continuous Algorithms, Part I
Convexity, logconcavity, and continuous algorithms

Kevin Tian

1 Introduction

Algorithm design has traditionally been studied through a discrete perspective. For example, classical textbooks [CLRS22] in large part focus on problems defined on naturally discrete domains.

Increasingly, however, modern research in algorithm design has benefitted from adopting a continuous perspective and using tools developed through the study of continuous mathematics. Sometimes, this benefit has come in discrete settings, e.g., *discrete optimization or sampling* problems, which ask to find the element x belonging to a finite set S which minimizes an objective $f(x)$, or to sample x proportional to a density $\mu(x)$. In such cases, it can often be helpful to consider a *continuous relaxation* of the discrete problem where the domain S is extended to a continuous superset S' containing it, and then a new *continuous optimization or sampling* problem is solved over S' (e.g., the Boolean cube $S = \{0,1\}^d$ can be relaxed to the hypercube $[0,1]^d$). The continuous solution is then converted to a desired element of S solving the original problem.

In other cases, often arising in the modern theory and practice of data science, the problem we are trying to solve is inherently continuous. For example, we could be performing parameter estimation in a statistical setting (e.g., learning the parameters of a generalized linear model, or estimating the top eigenvectors of a distribution's covariance). Additionally, outputs of continuous algorithms (e.g., samples from a distribution on $\mathbb{R}^d$) may have appealing properties for downstream use in a way discrete counterparts do not, giving guarantees such as robustness or data privacy.

As we hope to convey through this course, studying algorithms through a continuous lens is a rewarding experience for algorithm designers across a surprisingly diverse set of domains. Often, continuous methods provide frameworks for algorithm design which are quite distinct from their more traditional counterparts, offering new ways of exploiting structure latent in problems. What is particularly appealing about this continuous toolkit is the many synergies between its different components; in various situations, continuous methods provide unified, principled perspectives on algorithmic tools which may otherwise seem ad hoc. In this first lecture, we will touch upon two aspects of the continuous toolkit, centered around the analysis of convex and logconcave functions. We then give a first instructive example on how these analytical tools yield new powerful algorithmic primitives such as *convex programming*, for discrete and continuous problems alike.

2 Convexity

Convex analysis is the first central tool we encounter, and will be used throughout the course to develop algorithms. We begin with two definitions of convexity.

Definition 1 (Convex set)**.** *We say a set $\mathcal{X} \subseteq \mathbb{R}^d$ is* convex *if for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ the line segment between $\mathbf{x}$ and $\mathbf{x}'$ lies in $\mathcal{X}$, i.e., $(1-\lambda)\mathbf{x} + \lambda\mathbf{x}' \in \mathcal{X}$ for all $\lambda \in [0,1]$.*

Correspondingly, for any $\lambda \in [0,1]$, we call $(1-\lambda)\mathbf{x} + \lambda\mathbf{x}'$ a *convex combination* of $\mathbf{x}$ and $\mathbf{x}'$.

Definition 2 (Convex function)**.** *We say a function $f : \mathcal{X} \to \mathbb{R}$ is* convex *if $\mathcal{X}$ is convex and for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ the linear interpolation of $f(\mathbf{x})$ and $f(\mathbf{x}')$ overestimates the function on the line segment between $\mathbf{x}$ and $\mathbf{x}'$, i.e., $f((1-\lambda)\mathbf{x} + \lambda\mathbf{x}') \le (1-\lambda)f(\mathbf{x}) + \lambda f(\mathbf{x}')$ for all $\lambda \in [0,1]$.*

These two definitions are related as follows. Let $\chi_S(\mathbf{x})$ be the 0-∞ indicator of S, i.e., $\chi_S(\mathbf{x}) = 0$ if $\mathbf{x} \in S$ and $\chi_S(\mathbf{x}) = \infty$ otherwise. It is simple to verify that χ_S is a convex function iff S is a

convex set. Conversely, define the *epigraph* of $f : \mathcal{X} \to \mathbb{R}$ by

$$\text{epi}(f) := \{(\mathbf{x}, t) \in \mathcal{X} \times \mathbb{R} \mid t \geq f(\mathbf{x})\}.$$

We can also check that f is a convex function iff $\text{epi}(f)$ is a convex set.

Sometimes, Definition 2 in rephrased in terms of first-order approximations, in the following way.

Lemma 1. *Let $f : \mathcal{X} \to \mathbb{R}$ be differentiable. Then f is convex iff for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$,*

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle . \tag{1}$$

Proof. To show one direction, applying (1) at $\mathbf{x} \leftarrow \mathbf{x}_\lambda := (1-\lambda)\mathbf{x} + \lambda\mathbf{x}'$ and $\mathbf{x}' \leftarrow \mathbf{x}, \mathbf{x}'$ yields

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{x}_\lambda) + \langle \nabla f(\mathbf{x}_\lambda), \mathbf{x} - \mathbf{x}_\lambda \rangle = f(\mathbf{x}_\lambda) + \lambda \langle \nabla f(\mathbf{x}_\lambda), \mathbf{x} - \mathbf{x}' \rangle , \\ f(\mathbf{x}') &\geq f(\mathbf{x}_\lambda) + \langle \nabla f(\mathbf{x}_\lambda), \mathbf{x}' - \mathbf{x}_\lambda \rangle = f(\mathbf{x}_\lambda) + (1-\lambda) \langle \nabla f(\mathbf{x}_\lambda), \mathbf{x}' - \mathbf{x} \rangle , \end{aligned}$$

proving convexity of f upon linearly combining the two equations above. To show the other,

$$\begin{aligned} f(\mathbf{x}') &\geq \lim_{\lambda \to 0} \frac{f((1-\lambda)\mathbf{x} + \lambda\mathbf{x}') - (1-\lambda) f(\mathbf{x})}{\lambda} \\ &= f(\mathbf{x}) + \lim_{\lambda \to 0} \frac{f(\mathbf{x} + \lambda(\mathbf{x}' - \mathbf{x})) - f(\mathbf{x})}{\lambda} = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle . \end{aligned} \tag{2}$$

The only inequality used that the definition of f being convex holds pointwise over λ. □

Observe that the right-hand side of (1) is the first-order Taylor expansion of f about $\mathbf{x}$; Lemma 1 simply states that this first-order approximation underestimates f everywhere.

One of the basic reasons that convexity is useful is that convex functions have well-behaved minimizers. For example, we will frequently apply the following *first-order optimality condition*.

Lemma 2 (First-order optimality)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be differentiable and convex. Then*

$$\mathbf{x}^\star \in \text{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \iff \langle \nabla f(\mathbf{x}^\star), \mathbf{x}^\star - \mathbf{x} \rangle \leq 0, \text{ for all } \mathbf{x} \in \mathcal{X}.$$

Proof. To show one direction, if $\langle \nabla f(\mathbf{x}^\star), \mathbf{x}^\star - \mathbf{x} \rangle \leq 0$, directly applying (1) shows that

$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) + \langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star \rangle \geq f(\mathbf{x}^\star) \text{ for all } \mathbf{x} \in \mathcal{X}.$$

To show the other direction, let $\mathbf{x}^\star \in \text{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$, and suppose that $\langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star \rangle < 0$ for some $\mathbf{x} \in \mathcal{X}$, for the sake of contradiction. Then consider $\mathbf{x}_\lambda := (1-\lambda)\mathbf{x}^\star + \lambda\mathbf{x}$ and define $\phi(\lambda) := f(\mathbf{x}_\lambda)$. The derivation in (2) shows that $\phi'(0) = \langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star \rangle < 0$, so for small enough λ, we have $f(\mathbf{x}_\lambda) = \phi(\lambda) < \phi(0) = f(\mathbf{x}^\star)$, a contradiction to the optimality of $\mathbf{x}^\star$. □

Lemma 2 gives us a way of certifying optimality of a point $\mathbf{x}^\star$, by establishing $\langle \nabla f(\mathbf{x}^\star), \mathbf{x}^\star - \mathbf{x} \rangle \leq 0$ for all $\mathbf{x} \in \mathcal{X}$; notice this is certainly true if $\nabla f(\mathbf{x}^\star)) = \mathbf{0}_d$. Moreover, the second half of Lemma 2 already suggests a natural iterative approach to minimize convex functions. Whenever $\mathbf{x} \in \mathcal{X}$ is not a minimizer of f, there is a direction violating the first-order optimality condition, and moving a sufficiently short distance along this direction decreases the function value. We will expand upon this intuition and provide quantitative guarantees for it over the next few lectures.

The following characterization of the minima and maxima of convex functions is also often helpful.

Lemma 3. *Let $f : \mathcal{X} \to \mathbb{R}$ be convex, and let f have minimizer set $\mathcal{X}^\star := \text{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$ and maximizer set $\mathcal{X}^+ := \text{argmax}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$. Then $\mathcal{X}^\star$ is convex, and if $\mathcal{X}^+ \neq \emptyset$, it either contains a boundary point of $\mathcal{X}$ (i.e., $\mathbf{x}^+ \in \mathcal{X}$ such that $\nexists \mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $\lambda \in (0,1)$ with $\mathbf{x}^+ = (1-\lambda)\mathbf{x} + \lambda\mathbf{x}'$) or $\mathcal{X}$ has no boundary points. If f is strictly convex (i.e., Definition 2 holds with strict inequality), $\mathcal{X}^\star$ is a singleton and $\mathcal{X}^+$ only contains boundary points if it is nonempty.*

Proof. We begin with the claims about $\mathcal{X}^\star$. Suppose f has minimizers $\mathbf{x} \neq \mathbf{x}'$ (else there is nothing to prove), and let $f(\mathbf{x}) = f(\mathbf{x}') = f^\star$. For $\lambda \in [0,1]$ and $\mathbf{x}_\lambda := (1-\lambda)\mathbf{x}+\lambda\mathbf{x}'$, we have $f(\mathbf{x}_\lambda) \leq f^\star$ by convexity of f, so $\mathbf{x}_\lambda \in \mathcal{X}^\star$, proving convexity of $\mathcal{X}^\star$. If f is strictly convex and $f(\mathbf{x}) = f(\mathbf{x}') = f^\star$ for $\mathbf{x} \neq \mathbf{x}'$, strict convexity yields $\mathbf{x}_\lambda$ with $f(\mathbf{x}_\lambda) < f^\star$ for any $\lambda \in [0,1]$, a contradiction.

Next, we prove the claims about $\mathcal{X}^+$. If there is an interior point $\mathbf{x}^+ \in \mathcal{X}^+$ attaining the maximum function value $f(\mathbf{x}^+) = f^+$, choose any other $\mathbf{x} \in \mathcal{X}$, and note that by definition of interior points, the ray from $\mathbf{x}$ to $\mathbf{x}^+$ passes through another point $\mathbf{x}' \in \mathcal{X}$ extending beyond $\mathbf{x}^+$, so that $\mathbf{x}^+ = (1-\lambda)\mathbf{x}+\lambda\mathbf{x}'$ for $\lambda \in (0,1)$. Therefore, $f^+ = f(\mathbf{x}^+) \leq (1-\lambda)f(\mathbf{x})+\lambda f(\mathbf{x}')$, so $f(\mathbf{x}) = f^+$ and $\mathbf{x}$ is also a maximizer. Because $\mathbf{x}$ was arbitrary, f is a constant function, so if $\mathcal{X}$ has any boundary points they are in $\mathcal{X}^+$. If f is strictly convex, this rules out interior point maximizers. □

Let us now give an example of a type of convex function we will frequently encounter.

Lemma 4. *Let $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ be a seminorm on $\mathbb{R}^d$,[1] i.e., it satisfies the following properties.*

1. *Triangle inequality: $\|\mathbf{x}+\mathbf{x}'\| \leq \|\mathbf{x}\| + \|\mathbf{x}'\|$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$.*
2. *Absolute homogeneity: $\|t\mathbf{x}\| = |t| \, \|\mathbf{x}\|$ for all $t \in \mathbb{R}$.*

Then $\|\cdot\|$ is a convex function.

Proof. By the triangle inequality and absolute homogeneity respectively, we have the desired

$$\|(1-\lambda)\mathbf{x}+\lambda\mathbf{x}'\| \leq \|(1-\lambda)\mathbf{x}\| + \|\lambda\mathbf{x}'\| = (1-\lambda)\|\mathbf{x}\| + \lambda\|\mathbf{x}'\| \text{ for all } \lambda \in [0,1].$$

□

We next give a first application showcasing the utility of Lemmas 2 and 3.

Corollary 1. *Let $S \subset \mathbb{R}^d$ be compact and convex, and suppose $\mathbf{x}_0 \notin S$. There is a* separating hyperplane *$\mathbf{g} \in \mathbb{R}^d$, such that $\mathbf{g} \neq \mathbf{0}_d$ and $\mathbf{g}^\top\mathbf{x}_0 > \mathbf{g}^\top\mathbf{x}$ for all $\mathbf{x} \in S$. Moreover, if $\mathbf{x}_0 \in S$ is a boundary point, there is a* supporting hyperplane *$\mathbf{g} \neq \mathbf{0}_d \in \mathbb{R}^d$ with $\mathbf{g}^\top\mathbf{x}_0 \geq \mathbf{g}^\top\mathbf{x}$ for all $\mathbf{x} \in S$.*

Proof. For the first claim, let $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}-\mathbf{x}_0\|_2^2$; it is straightforward to check f is strictly convex,[2] and $\nabla f(\mathbf{x}) = \mathbf{x}-\mathbf{x}_0$. Let $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x}\in S} f(\mathbf{x})$, which exists since we are minimizing a continuous function over a compact set; Lemma 3 guarantees $\mathbf{x}^\star$ is unique. Then, by Lemma 2,

$$\langle \nabla f(\mathbf{x}^\star), \mathbf{x}^\star - \mathbf{x}\rangle = \langle \mathbf{x}^\star - \mathbf{x}_0, \mathbf{x}^\star - \mathbf{x}\rangle \leq 0 \text{ for all } \mathbf{x} \in S.$$

Let $\mathbf{g} := \mathbf{x}_0 - \mathbf{x}^\star \neq \mathbf{0}^d$, since $\mathbf{x}_0 \notin S$. We then rearrange the above, showing the desired

$$\mathbf{g}^\top\mathbf{x} \leq \mathbf{g}^\top\mathbf{x}^\star = \mathbf{g}^\top\mathbf{x}_0 + \mathbf{g}^\top(\mathbf{x}^\star - \mathbf{x}_0) = \mathbf{g}^\top\mathbf{x}_0 - \|\mathbf{g}\|_2^2 < \mathbf{g}^\top\mathbf{x}_0 \text{ for all } \mathbf{x} \in S.$$

For the second claim, we can take a convergent subsequence $\{\mathbf{x}_i\}_{i\geq 1} \subset \mathbb{R}^d \backslash S$ approaching $\mathbf{x}_0$, which come with supporting hyperplanes $\{\mathbf{g}_i\}_{i\geq 1} \in \mathbb{R}^d$ which are, without loss of generality, unit length. Taking $\mathbf{g}$ to be the limit of any convergent subsequence of the $\{\mathbf{g}_i\}_{i\geq 1}$, we have $\mathbf{g}^\top(\mathbf{x}_0 - \mathbf{x}) = \lim_{i\to\infty} \mathbf{g}_i^\top(\mathbf{x}_i - \mathbf{x}) \geq 0$ for all $\mathbf{x} \in S$, which yields the claim upon rearranging. □

We pause to address an important point: the assumption of differentiability in Lemmas 1 and 2. Notice that the definition of convexity (Definition 2) and our other results in this section do not use differentiability; indeed, there are convex functions (e.g., $f(x) = |x|$) which are not differentiable everywhere. Moreover, the definition of convexity does not even rule out discontinuous functions. Fortunately, even in such scenarios convex functions admit the following proxy for a derivative.

Definition 3 (Subgradient)**.** *Let $f : \mathcal{X} \to \mathbb{R}$. We say $\mathbf{g}$ is a* subgradient *of f at $\mathbf{x} \in \mathcal{X}$ if*

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{x}' - \mathbf{x}\rangle \text{ for all } \mathbf{x}' \in \mathcal{X}.$$

We denote the set of subgradients of f at $\mathbf{x}$ by $\partial f(\mathbf{x})$.

[1] Seminorms are norms without the positive definiteness restriction that only $\mathbf{x} = \mathbf{0}_d$ has $\|\mathbf{x}\| = 0$.

[2] We will develop several ways to verify convexity more easily in later lectures, but for now note that strict convexity of f follows from a direct expansion of Definition 2 and completing the square.

Comparing to (1), it is clear that if f is convex, ∇f is a subgradient everywhere it is defined. Interestingly, convex functions admit subgradients almost everywhere they are defined.

Lemma 5. *Let $f : \mathcal{X} \to \mathbb{R}$ be convex, and assume $\mathcal{X} \subseteq \mathbb{R}^d$. For all $\mathbf{x} \in \text{relint}(\mathcal{X})$, the relative interior of $\mathcal{X}$,[3] $\partial f(\mathbf{x})$ is nonempty.*

Proof. Since $(\mathbf{x}, f(\mathbf{x}))$ lies on the boundary of $\text{epi}(f)$, Corollary 1 gives $(\mathbf{a}, b) \neq \mathbf{0}_{d+1} \in \mathbb{R}^d \times \mathbb{R}$ such that for all $(\mathbf{y}, t) \in \text{epi}(f)$, $\mathbf{a}^\top \mathbf{x} + bf(\mathbf{x}) \geq \mathbf{a}^\top \mathbf{y} + bt$. We may assume without loss that $\mathbf{a}$ is in the minimal subspace containing $\mathcal{X}$. Since t can be arbitrarily large, this implies $b \leq 0$.

We claim $b \neq 0$. If $\mathbf{a} = \mathbf{0}_d$, then indeed $b \neq 0$ since $(\mathbf{a}, b) \neq \mathbf{0}_{d+1}$. Otherwise, as $\mathbf{x}$ lies in the relative interior of $\mathcal{X}$, choosing $\mathbf{y} = \mathbf{x} + \epsilon \mathbf{a} \in \mathcal{X}$ for sufficiently small $\epsilon > 0$ also gives $b \neq 0$, else $\mathbf{a}^\top \mathbf{x} < \mathbf{a}^\top \mathbf{y}$ would be a contradiction. Finally, since $(\mathbf{x}', f(\mathbf{x}')) \in \text{epi}(f)$, we have for $\mathbf{g} = -\frac{1}{b} \cdot \mathbf{a}$,

$$\mathbf{a}^\top \mathbf{x} + bf(\mathbf{x}) \geq \mathbf{a}^\top \mathbf{x}' + bf(\mathbf{x}') \implies f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{x}' - \mathbf{x} \rangle \implies \mathbf{g} \in \partial f(\mathbf{x}).$$

□

While pathological examples do exist for convex functions, generic properties such as Lemma 5 relying on fairly minimal assumptions provide convex functions a great deal of regularity, and allow us to design general-purpose algorithms. For the remainder of the course, to avoid pathological examples, we adopt the following assumptions anytime we discuss a convex function $f : \mathcal{X} \to \mathbb{R}$, except when otherwise stated, which simplifies much of our downstream development.

1. f is *closed*, i.e., its epigraph $\text{epi}(f)$ is closed. It is straightforward to verify this implies f is continuous within the relative interior of $\mathcal{X}$.[4] Two common examples of closed functions are functions which are finite on a closed set $\mathcal{X}$, and functions of Legendre type. A function is of Legendre type if it is differentiable everywhere in $\mathcal{X} := \text{int}(\{\mathbf{x} \in \mathbb{R}^d \mid f(\mathbf{x}) < \infty\}) \neq \emptyset$, and $f, \nabla f \to \infty$ as x approaches the boundary of $\mathcal{X}$. The fact that $f \to \infty$ as $\mathbf{x}$ approaches the boundary prevents existence of a limit point of $\text{epi}(f)$ that is not contained in $\text{epi}(f)$.
2. f is *proper*, i.e., it takes $\mathbb{R}^d$ to values in $\mathbb{R} \cup \{\infty\}$, and is finite on $\mathcal{X} \neq \emptyset$. We often overload f with its proper extension, defining $f(\mathbf{x}) = \infty$ for any $\mathbf{x} \notin \mathcal{X}$, when $\mathcal{X} \subset \mathbb{R}^d$. Correspondingly, when we say $f : \mathcal{X} \to \mathbb{R}$ is convex, we imply $\mathcal{X}$ is the set where f is finite.

As we will see, Section 3 develops a general-purpose algorithm which minimizes convex functions, relying on convexity only through Lemma 3, Corollary 1, and Lemma 5, highlighting how fundamental these results are. Indeed, Section 3 demonstrates that convex functions are appealing from an algorithmic perspective, as convexity of f implies an efficient algorithm for optimizing f.

We mention that convexity is a property of significant relevance in applications.

1. Linear functions and polytopes (intersections of halfspaces) are both convex, and hence convex optimization applies to the ubiquitous problem of linear programming (linear optimization over a polytope). Problems which can be written as linear programs are widespread, e.g., minimum-cost flow and its relatives, resource allocation, and various scheduling problems.
2. Common objectives in statistics and machine learning, such as linear regression, logistic regression, support vector machines, and regularizers such as the Lasso and ElasticNet, are all convex. Other problems are modeled with convex relaxations, such as the ELBO loss in variational Bayesian methods, or semidefinite programs (a generalization of linear programs).
3. Discrete optimization problems defined over subsets of a base set may also be amenable to convex optimization algorithms. For example, a submodular function is defined over $\{0, 1\}^S$ for a discrete set S, but admits a continuous relaxation (the "Lovasz extension") which is convex. This relaxation has the useful properties that we can efficiently compute subgradients of it, and by convexity, the minimizer of the relaxation is an extremal point and hence an element of $\{0, 1\}^S$. Submodularity is a property which captures the notion of "diminishing marginal returns," and often models problems where diversity is a target.

[3] Recall the interior of a set S is all points in S with an open neighborhood in S. In settings where $S \subseteq \mathbb{R}^d$ but S is not full-dimensional, the relative interior of S is the interior of S within the smallest subspace containing it.

[4] One may hope that closedness implies Lemma 5 can be modified to hold true everywhere on $\mathcal{X}$, not just the interior. However, the example $f(x) = -\sqrt{1 - x^2}$ for $x \in [-1, 1]$ and $f(x) = \infty$ elsewhere dashes these hopes, as no subgradient exists at $x = \pm 1$. Nonetheless, closedness is a useful assumption in other pervasive situations.

Our study of convex analysis will prove fruitful beyond convex optimization; throughout the course, we highlight several examples of functions which are nonconvex, yet nonetheless admit efficient optimization algorithms. The development of these structured *nonconvex optimization* algorithms will draw heavily upon our convex analysis tools, adding further merit to the study of this theme.

3 Cutting-plane methods

In this section, we provide an application of the facts shown in Section 2. Specifically, we will establish the following remarkable theorem by designing an algorithm.

Theorem 1 (Polynomial-time convex optimization)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be convex for $\mathcal{X} \subset \mathbb{R}^d$, and assume f has an additive range[5] bounded by $\mathrm{poly}(d)$. There is an algorithm which uses $O(d \log \frac{d}{\epsilon})$ queries to a value and subgradient oracle for f, and $\mathrm{poly}(d, \log \frac{1}{\epsilon})$ additional time, such that with high probability,[6] the algorithm returns $\mathbf{x}$ satisfying $f(\mathbf{x}) \leq \min_{\mathbf{x}^\star \in \mathcal{X}} f(\mathbf{x}^\star) + \epsilon$.*

To build up to Theorem 1, we begin with a statement of a conceptual framework for algorithm design, phrased as a game. In it, a player (algorithm designer) Alice is attempting to end the game, and an adversary Bob is trying to make Alice's job as difficult as possible while playing by the rules. There are two ways the game can end: either Alice finds a point in a hidden set $S^\star$, or Alice sufficiently reduces the volume of a superset of $S^\star$. We now formally define this game.

Definition 4 (Cutting-plane game)**.** *Consider the following game between Bob, who holds compact, convex $S^\star \subset \mathbb{R}^d$, and Alice, who holds $S_0 \supseteq S^\star$, starting from $t = 0$ and parameterized by $V_{\min} > 0$.*

1. *On turn t, if $\mathrm{Vol}(S_t) < V_{\min}$, the game ends. Else, Alice chooses $\mathbf{x}_t \in S_t$.*
2. *If $\mathbf{x}_t \in S^\star$, the game ends. Else, Bob chooses $\mathbf{g}_t \in \mathbb{R}^d$ such that $\mathbf{g}_t^\top \mathbf{x}_t > \mathbf{g}_t^\top \mathbf{x}$ for all $\mathbf{x} \in S^\star$. Note that such $\mathbf{g}_t$ exists by Corollary 1, but is not necessarily unique.*
3. *Alice updates S_{t+1} to be any superset of $S_t \cap H_t$ where $H_t := \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{g}_t^\top \mathbf{x} < \mathbf{g}_t^\top \mathbf{x}_t\}$, and the game advances to turn $t + 1$.*

Observe that the definitions of Steps 2 and 3 imply the invariant $S_t \supseteq S^\star$ for all iterations t where the game is played. However, there is substantial freedom in how the two players play the game: in particular, how should Alice choose $\mathbf{x}_t \in S_t$ and, upon observing $\mathbf{g}_t$, update S_t to S_{t+1}? Conversely, what choice of $\mathbf{g}_t$ would make Alice's job as hard as possible? We will shortly give an instantiation of Alice's strategy, which rapidly terminates the game regardless of Bob's strategy.

One significant reason for studying the cutting-plane game is because it naturally captures convex optimization under a first-order access model as an application. The following observation also explains why we give Alice the win condition of decreasing the volume of a set sufficiently: the remaining volume scales with the approximation error for solving the optimization problem.

Lemma 6. *Let $f : \mathcal{X} \to \mathbb{R}$ be convex for $\mathcal{X} \subset \mathbb{R}^d$, and suppose f has minimizer set $\mathcal{X}^\star := \mathrm{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$.[7] Suppose we play the cutting-plane game (Definition 4) initialized from $S_0 \leftarrow \mathcal{X}$, and let $\alpha := \frac{V_{\min}}{\mathrm{Vol}(S_0)} \in (0, 1)$. Further, suppose Alice always chooses $\mathbf{x}_t \in \mathrm{relint}(S_t)$, and Bob (who holds $S^\star \leftarrow \mathcal{X}^\star$) plays by ending the game if $\mathbf{0}_d \in \partial f(\mathbf{x}_t)$, and returning $\mathbf{g}_t \neq \mathbf{0}_d \in \partial f(\mathbf{x}_t)$ otherwise. If the game terminates in T iterations, letting $f^\star := \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$, we have*

$$\min_{t \in [T]} f(\mathbf{x}_t) \leq f^\star + \alpha^{\frac{1}{d}} \left(\max_{\mathbf{z} \in \mathcal{X}} f(\mathbf{z}) - f^\star \right).$$

Proof. We first verify that Bob's implementation, which returns $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$, is valid for Step 2 in Definition 4 (note that $\partial f(\mathbf{x}_t) \neq \emptyset$ by Lemma 5). First, checking whether $\mathbf{0}_d \in \partial f(\mathbf{x}_t)$ is equivalent to checking whether $\mathbf{x}_t \in \mathcal{X}^\star$, by Lemma 7. If $\mathbf{0}_d \notin \partial f(\mathbf{x}_t)$, we observe that for $\mathbf{x}^\star \in \mathcal{X}^\star$:

$$0 > f(\mathbf{x}^\star) - f(\mathbf{x}_t) \geq \langle \mathbf{g}_t, \mathbf{x}^\star - \mathbf{x}_t \rangle . \tag{3}$$

[5] That is, $\max_{x \in \mathcal{X}} f(x) - \min_{x \in \mathcal{X}} f(x)$.

[6] We will be more precise with the dependence of runtimes, etc. on failure probabilities when formally proving guarantees on randomized algorithms throughout the course. For now, "with high probability" implies that all complexities depend polylogarithmically on the inverse failure probability.

[7] In cases when f is unconstrained (i.e., $\mathcal{X} = \mathbb{R}^d$), we assume that we have knowledge of $\mathcal{X}$ containing $\mathcal{X}^\star$.

The first inequality used that $\mathbf{x}_t \notin \mathcal{X}^\star$. Next, notice that

$$\mathbf{x} \notin H_t \implies f(\mathbf{x}) \geq f(\mathbf{x}_t) + \mathbf{g}_t^\top(\mathbf{x} - \mathbf{x}_t) \geq f(\mathbf{x}_t), \tag{4}$$

where the first inequality used $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ and the second used $\mathbf{x} \notin H_t$. Finally, consider the set

$$S_\alpha := \left\{\mathbf{x} \mid \mathbf{x} = (1-\alpha^{\frac{1}{d}})\mathbf{x}^\star + \alpha^{\frac{1}{d}}\mathbf{z}, \text{ for } \mathbf{z} \in \mathcal{X}\right\} = \{(1-\alpha^{\frac{1}{d}})\mathbf{x}^\star\} \oplus \alpha^{\frac{1}{d}}\mathcal{X}.$$

We defined S_α so that $\mathrm{Vol}(S_\alpha) = \mathrm{Vol}(\alpha^{\frac{1}{d}}\mathcal{X}) = \alpha\mathrm{Vol}(S_0) = V_{\min}$. Now, the game can either end because Alice has found $\mathbf{x}_T \in \mathcal{X}^\star$, for which the conclusion is clearly true, or because $\mathrm{Vol}(S_T) < V_{\min}$. In the latter case, there is a point $\mathbf{x} \in S_\alpha \setminus S_T$, such that $\mathbf{x} = (1-\alpha^{\frac{1}{d}})\mathbf{x}^\star + \alpha^{\frac{1}{d}}\mathbf{z}$. Notice that by the definition of Step 3, the only way $\mathbf{x} \notin S_T$ is if $\mathbf{x} \notin H_t$ for some $t \in [T]$. On that iteration t, applying (4) shows the desired claim for $\mathbf{x}_t$, as

$$f(\mathbf{x}_t) \leq f(\mathbf{x}) \leq (1-\alpha^{\frac{1}{d}})f^\star + \alpha^{\frac{1}{d}}f(\mathbf{z}) \leq f^\star + \alpha^{\frac{1}{d}}\left(\max_{\mathbf{z}\in\mathcal{X}} f(\mathbf{z}) - f^\star\right).$$

□

In Lemma 6, we used the following simple observation.

Lemma 7. *Let $f : \mathcal{X} \to \mathbb{R}$, and let $\mathbf{x}^\star \in \mathcal{X}$. Then $\mathbf{x}^\star \in \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \iff \mathbf{0}_d \in \partial f(\mathbf{x}^\star)$.*

Proof. This follows from the sequence of equivalences:

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{X} \iff f(\mathbf{x}^\star) + \mathbf{0}_d^\top(\mathbf{x} - \mathbf{x}^\star) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{X} \iff \mathbf{0}_d \in \partial f(\mathbf{x}^\star).$$

□

Lemma 6 gives a powerful, general-purpose algorithm template for convex optimization in the *oracle model*, assuming that we can design a good strategy for Alice to terminate the cutting-plane game. Specifically, assume access to a value oracle and a subgradient oracle for f, defined in Definition 5. If we assume $f(\mathbf{z}) - f^\star \leq \Delta$ for all $\mathbf{z} \in \mathcal{X}$, and we wish to produce a point $\mathbf{x} \in \mathcal{X}$ with $f(\mathbf{x}) \leq f^\star + \epsilon$, it suffices to take $\alpha = (\frac{\epsilon}{\Delta})^d$, and call Lemma 6. We implement Bob's strategy in the cutting-plane game using the subgradient oracle, and return the iterate we encounter with the smallest function value, by querying all of our iterates using the value oracle.

Definition 5 (Value and subgradient oracle)**.** *We say $\mathcal{O}$ is a* value oracle *for $f : \mathcal{X} \to \mathbb{R}^d$ if when queried at $\mathbf{x} \in \mathbb{R}^d$, it returns $f(\mathbf{x})$ if $\mathbf{x} \in \mathcal{X}$ and ∞ otherwise. We say $\mathcal{O}$ is a* subgradient oracle *for $f : \mathcal{X} \to \mathbb{R}^d$ if when queried at $\mathbf{x} \in \mathbb{R}^d$, it returns an element of $\partial f(\mathbf{x})$ if it exists (set to $\mathbf{0}_d$ by default if $\mathbf{0}_d \in \partial f(\mathbf{x})$), and otherwise returns nothing. When f is differentiable, we also call any subgradient oracle for f a* gradient oracle*, which uniquely returns $\nabla f(\mathbf{x})$ when queried at $\mathbf{x}$.*

Typically, it is reasonable to assume that $\frac{\Delta}{\epsilon}$ is polynomially bounded in d, in which case $\frac{1}{\alpha} = d^{O(d)}$. In other words, we want to implement Alice's strategy in a way which reduces the volume of S_0 by a $d^{O(d)}$ factor, in few iterations. We describe one such strategy, relying on the following theorem.

Theorem 2 (Grünbaum)**.** *Let $S \subseteq \mathbb{R}^d$ be convex, and let*

$$\bar{\mathbf{x}}_S := \frac{1}{\mathrm{Vol}(S)} \int \mathbf{x} \exp(-\chi_S(\mathbf{x}))\mathrm{d}\mathbf{x} \tag{5}$$

denote the center of gravity of S. Then any halfspace $H = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{v}^\top\mathbf{x} \leq \mathbf{v}^\top\bar{\mathbf{x}}_S\}$ whose defining halfplane passes through $\bar{\mathbf{x}}_S$ satisfies $\mathrm{Vol}(S \cap H) \geq \frac{1}{e}\mathrm{Vol}(S)$.

We prove Theorem 2, first shown by [Gru60], in Section 5. Here, we observe that it immediately yields a strategy for Alice, known as the "center of gravity" method in the literature. If Alice simply maintains $S_{t+1} = S_t \cap H_t$, and chooses $\mathbf{x}_t = \bar{\mathbf{x}}_{S_t}$ (i.e., the center of gravity of S_t) in Step 1 each turn, then any halfspace will guarantee $\mathrm{Vol}(S_{t+1}) \leq (1-\frac{1}{e})\mathrm{Vol}(S_t)$ by Theorem 2. Therefore, in $T = O(d \log d)$ iterations we can ensure the volume of S_0 is reduced by a $d^{O(d)}$ factor.

Remark 1. *It is not a priori clear how to implement each step of the center of gravity method efficiently, because it requires computing the center of gravity $\bar{\mathbf{x}}_{S_t}$ in each iteration t. As we discuss in Theorem 3 in Section 4, we can produce approximate samples from the uniform distribution on S_t, and as shown in [BV04], averaging a polynomial number of these approximate samples suffices for an approximate variant of Theorem 2 to hold with high probability, i.e., where the constant $\frac{1}{e}$ is replaced with a smaller constant. Intuitively, we are computing an approximate center of gravity.*

In light of Remark 1 and the framework above, we have proven Theorem 1 (which relies on our polynomial-time approximate sampler, in turn stated later in Section 4 as Theorem 3).

In fact, using a quantitative variant of Alexandrov's theorem (which states that convex functions are differentiable almost everywhere), [LSV18] shows that Theorem 1 can be improved to not even require a subgradient oracle by simulating subgradient computations using a value oracle and a finite-differences method, losing an $\approx d$ factor in the value oracle query complexity.

Our proof of Theorem 1 is surprising because it essentially shows that binary search is possible to do in polynomial time, even in high dimensions. Given an initial volume, which is typically exponentially-sized in the dimension, we can quickly pin down a small set containing the minimizer by repeatedly querying a subgradient oracle. More generally, this center of gravity method gives a way of playing the cutting-plane game against an adversarial separation oracle.

Remark 2. *The center of gravity method is just one strategy for Alice to play the cutting-plane game. There are other algorithms, collectively "cutting-plane methods," which provide guarantees for the cutting-plane game trading off the number of iterations before termination, and the additional computation required by Alice. As we will see, the center-of-gravity method attains an optimal iteration count, but requires a very expensive (though still polynomial-time) implementation.*

For example, because S_t can become very complicated over time, one may elect to use a cheaper superset to approximate it. The ellipsoid method does this by maintaining an approximating ellipsoid instead [Kha80]. This cutting-plane method is relatively cheap to implement (requiring $\approx d^2$ time per iteration to maintain a matrix defining an ellipse), but loses a factor of d in the number of iterations over the center of gravity method because of a worse volume decrease guarantee.

A line of work [Vai96, LSW15, JLSW20] has developed implementations which have gradually improved the cost of cutting-plane methods to match the iteration complexity of the center of gravity method, and require $O(d^2)$ additional computation per iteration matching the ellipsoid method, which is intuitively necessary to maintain S_t by updating its constraint matrix.

Theorem 1, and its efficient implementation (as given by Theorem 3 and the recent works mentioned in Remark 2), represent a powerful way to establish the polynomial-time tractability of a convex optimization problem, simply by appealing to convexity. However, the actual computational overhead of this general-purpose tool ($\Omega(d^3)$ in the worst case) can still be highly superlinear for many natural problems, which admit $o(d^3)$-sized descriptions. Throughout the first part of the course, we give a variety of alternative *structured optimization* algorithms. In appropriate convex optimization settings, these improved algorithms allow us to beat the black box of calling Theorem 1 and obtain improved runtimes by catering to problem-specific structure.

4 Logconcavity

We introduce our next topic of study, logconcave functions, in this section. Analogously to how convexity is a sign of tractability for continuous optimization problems (and convex analysis is useful in broader, potentially nonconvex, settings), the analysis of logconcave functions is a very useful tool when designing statistical algorithms in continuous settings. For example, later lectures develop a sampling algorithm that applies generically to logconcave distributions, demonstrating that logconcavity is a sign of tractability for sampling problems. We begin with a definition.

Definition 6 (Logconcave function)**.** *We say a function $\mu : \mathcal{X} \to \mathbb{R}_{>0}$ is* logconcave *if $\mathcal{X}$ is convex and for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and all $\lambda \in [0, 1]$, $\mu((1-\lambda)\mathbf{x} + \lambda\mathbf{x}') \geq \mu(\mathbf{x})^{1-\lambda}\mu(\mathbf{x}')^{\lambda}$.*

To demystify Definition 6, taking a logarithm and negating shows for $f := -\log \mu$,

$$f((1-\lambda)\mathbf{x} + \lambda\mathbf{x}') \leq (1-\lambda)f(\mathbf{x}) + \lambda f(\mathbf{x}'),$$

i.e., $f : \mathcal{X} \to \mathbb{R}$ is convex. In accordance with our discussion of proper extensions of convex functions at the end of Section 2, we always assume that a logconcave function $\mu : \mathcal{X} \to \mathbb{R}_{>0}$ takes on the value 0 outside $\mathcal{X}$, which corresponds to $-\log \mu$ taking on the value ∞.

Intuitively, just as convexity of f prevents it from having disjoint sets of minimizers (Lemma 3), logconcavity of μ (when μ is a probability density) prevents it from having disjoint modes, allowing sampling algorithms to locally explore. Notice that logconcave functions are not necessarily probability densities (i.e., we may have $\int \mu(\mathbf{x})\mathrm{d}\mathbf{x} \neq 1$), but whenever μ is integrable, there is a normalizing constant $Z := \int \mu(\mathbf{x})\mathrm{d}\mathbf{x}$ such that $\frac{\mu}{Z}$ is a probability density, so we will often conflate a logconcave μ with the associated density $\propto \mu$. We call logconcave $\mu : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ a *logconcave density* if $\int \mu(\mathbf{x})\mathrm{d}\mathbf{x} = 1$. For intuition, a few canonical logconcave functions follow.

1. Let $\mu(\mathbf{x}) = \exp(-\frac{1}{2}\left\|\mathbf{x}\right\|_2^2)$ be the (unnormalized) standard multivariate Gaussian density on $\mathbb{R}^d$. We can verify μ is logconcave by observing that $f := -\log \mu = \frac{1}{2}\left\|\cdot\right\|_2^2$ is convex.
2. Let $\mu(\mathbf{x})$ be the 0-1 indicator of a convex set S, corresponding to the (unnormalized) uniform distribution on S. This μ is also logconcave since $f := -\log \mu = \chi_S$ is convex.

While it is straightforward to sample from the density $\propto \exp(-\frac{1}{2}\left\|\cdot\right\|_2^2)$ since it decomposes into independent coordinates, it is less obvious how to efficiently sample from the density $\propto \exp(-\chi_S)$ for potentially ill-behaved S. By using tools from logconcave analysis (and other continuous techniques), we develop an algorithm for solving a significant generalization of this uniform sampling problem later in the course, informally summarized in the following.

Theorem 3 (Polynomial-time logconcave sampling)**.** *Let $\mu : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a logconcave density, and let $\mu \propto \exp(-f)$ where $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex and $\mathrm{poly}(d)$-well-conditioned where it is finite.*[8] *There is an algorithm which uses $\mathrm{poly}(d, \log \frac{1}{\epsilon})$ queries to a value oracle for f, and $\mathrm{poly}(d, \log \frac{1}{\epsilon})$ additional time, to produce a sample within ϵ total variation distance of μ.*

Recall that Theorem 1 is a powerful general-purpose convex optimization primitive, but can be substantially improved in structured settings. Along with proving Theorem 3 in the second part of the course, we show how to design *structured sampling* algorithms which can substantially improve upon the runtimes of Theorem 3 when the density μ admits additional structure.

The most useful inequality in studying logconcave functions is the Prékopa-Leindler inequality [Pre73]. Indeed, many results in logconcave analysis may be rephrased as an application of it.

Theorem 4 (Prékopa-Leindler)**.** *Let $\lambda \in (0, 1)$, and let $f, g, h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ satisfy $h((1-\lambda)\mathbf{x} + \lambda\mathbf{x}') \geq f(\mathbf{x})^{1-\lambda} g(\mathbf{x}')^{\lambda}$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. Then*

$$\int h(\mathbf{x})\mathrm{d}\mathbf{x} \geq \left(\int f(\mathbf{x})\mathrm{d}\mathbf{x}\right)^{1-\lambda} \left(\int g(\mathbf{x})\mathrm{d}\mathbf{x}\right)^{\lambda}.$$

Theorem 4 is daunting, but is related to the much more interpretable Brunn-Minkowski inequality, which we present shortly. To gain intuition for Theorem 4 we first give two simple consequences.

Corollary 2. *Let $\mu, \mu' : \mathbb{R}^d \to \mathbb{R}$ be logconcave.*

1. *Let $S \subseteq [d]$ and let $\mu_S(\mathbf{x}_S) := \int_{\mathbb{R}^{[d]\setminus S}} \mu(\mathbf{x}_S, \mathbf{x}_{-S})\mathrm{d}\mathbf{x}_{-S}$ be the* marginal *on S, defined for $\mathbf{x}_S \in \mathbb{R}^S$. Then μ_S is logconcave, and if μ is a density, so is μ_S.*
2. *Let $(\mu * \mu')(\mathbf{x})$ be the* convolution *of μ and μ', i.e., $(\mu * \mu')(\mathbf{x}) = \int \mu(\mathbf{x} - \mathbf{y})\mu'(\mathbf{y})\mathrm{d}\mathbf{y}$. Then $\mu * \mu'$ is logconcave, and if μ, μ' are densities, so is $\mu * \mu'$.*

Proof. By integrating μ_S over $\mathbf{x}_S \in \mathbb{R}^S$, it is clear that μ_S is a density if we assume $\int_{\mathbb{R}^d} \mu(\mathbf{x})\mathrm{d}\mathbf{x} = 1$. To see logconcavity of μ_S, fix two points $\mathbf{x}_S, \mathbf{x}'_S \in \mathbb{R}^S$, and let $\lambda \in (0, 1)$. Since logconcavity of μ verifies the precondition of Theorem 4 with $f, g, h : \mathbb{R}^{[d]\setminus S} \to \mathbb{R}$ defined by

$$f(\mathbf{x}_{-S}) := \mu(\mathbf{x}_S, \mathbf{x}_{-S}),\ g(\mathbf{x}_{-S}) := \mu(\mathbf{x}'_S, \mathbf{x}_{-S}),\ \text{and}\ h(\mathbf{x}_{-S}) := \mu((1-\lambda)\mathbf{x}_S + \lambda\mathbf{x}'_S, \mathbf{x}_{-S}),$$

applying Theorem 4 with these functions proves logconcavity of μ_S.

[8] We will formally define well-conditionedness in a later lecture, which roughly is a measure of the regularity of f. The assumption that f is $\mathrm{poly}(d)$-well-conditioned is not restrictive in practice.

Similarly, if μ, μ' are densities, $(\mu * \mu')$ is clearly a density upon integrating over $\mathbf{x}$, since

$$\int (\mu * \mu')(\mathbf{x}) \mathrm{d}\mathbf{x} = \left(\int \mu(\mathbf{z}) \mathrm{d}\mathbf{z} \right) \left(\int \mu'(\mathbf{y}) \mathrm{d}\mathbf{y} \right) = 1,$$

where each $(\mathbf{z}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d$ is counted once on the left-hand side corresponding to $\mathbf{x} = \mathbf{z} + \mathbf{y}$. To prove that $\mu * \mu'$ is logconcave if both μ and μ' are, fix $\mathbf{x}, \mathbf{x}'$ and $\lambda \in (0,1)$, so we wish to show

$$\int \mu \left(\underbrace{(1-\lambda)\mathbf{x} + \lambda \mathbf{x}'}_{:= \mathbf{x}_\lambda} - \mathbf{z} \right) \mu'(\mathbf{z}) \, \mathrm{d}\mathbf{z} \geq \left(\int \mu(\mathbf{x} - \mathbf{z}) \mu'(\mathbf{z}) \mathrm{d}\mathbf{z} \right)^{1-\lambda} \left(\int \mu(\mathbf{x}' - \mathbf{z}) \mu'(\mathbf{z}) \right)^{\lambda}.$$

Define $f(\mathbf{z}) := \mu(\mathbf{x} - \mathbf{z})\mu'(\mathbf{z})$, $g(\mathbf{z}) := \mu(\mathbf{x}' - \mathbf{z})\mu'(\mathbf{z})$, and $h(\mathbf{z}) := \mu(\mathbf{x}_\lambda - \mathbf{z})\mu'(\mathbf{z})$, so that applying Theorem 4 to these functions immediately implies the above claim, if its precondition is met. Indeed, for any $\mathbf{z}, \mathbf{z}'$, defining $\mathbf{z}_\lambda := (1-\lambda)\mathbf{z} + \lambda \mathbf{z}'$, we have the desired

$$h(\mathbf{z}_\lambda) = \mu(\mathbf{x}_\lambda - \mathbf{z}_\lambda) \mu'(\mathbf{z}_\lambda) \geq \left(\mu(\mathbf{x} - \mathbf{z})^{1-\lambda} \mu(\mathbf{x}' - \mathbf{z}')^{\lambda} \right) \left(\mu'(\mathbf{z})^{1-\lambda} \mu'(\mathbf{z}')^{\lambda} \right) = f(\mathbf{z})^{1-\lambda} g(\mathbf{z}'),$$

where we applied logconcavity of μ and μ' in the only inequality. □

Theorem 4 and its consequences provide a powerful set of tools for analyzing probabilistic statements. For example, suppose we wish to show, for convex $S \subseteq \mathbb{R}^d$ symmetric about $\mathbf{0}_d$,

$$f(\mathbf{x}) := \Pr_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)} [\mathbf{z} \in S \oplus \{\mathbf{x}\}]$$

is maximized when $\mathbf{x} = \mathbf{0}_d$. In other words, we want to show that a random Gaussian vector in $\mathbb{R}^d$ falls in $S \oplus \{\mathbf{x}\}$ with the highest probability when $\mathbf{x} = \mathbf{0}_d$. While intuitively obvious, it is not a priori clear how one would show such a statement. Consider applying the second part of Corollary 2 with $\mu \propto \exp(-\frac{1}{2} \|\cdot\|_2^2)$ and $\mu' = \exp(-\chi_S(\cdot))$, where χ_S is the indicator function of S, and μ is normalized to be a density. This gives a logconcave function $\mu * \mu'$ which, when evaluated at $\mathbf{x} \in \mathbb{R}^d$, exactly corresponds to the probability $f(\mathbf{x})$ defined above:

$$(\mu * \mu')(\mathbf{x}) = \int \mu(\mathbf{z}) \mu'(\mathbf{x} - \mathbf{z}) \mathrm{d}\mathbf{z} = \int \mu(\mathbf{z}) \mathbb{I}_{\mathbf{x} - \mathbf{z} \in S} \mathrm{d}z = \int \mu(\mathbf{z}) \mathbb{I}_{\mathbf{z} \in S \oplus \{\mathbf{x}\}} \mathrm{d}\mathbf{z} = f(\mathbf{x}).$$

Moreover, f is clearly symmetric about $\mathbf{0}_d$. Since symmetric, convex functions are minimized at $\mathbf{0}_d$, symmetric, logconcave functions are also maximized there, giving the conclusion.

We next sketch how to prove Theorem 4 by starting with the seemingly-simpler Brunn-Minkowski inequality. In fact, Theorem 4 implies Theorem 5, so we prove this first.

Theorem 5 (Brunn-Minkowski)**.** *Let A, B be compact and d-dimensional.*[9] *Then* $\mathrm{Vol}(A \oplus B)^{\frac{1}{d}} \geq (\mathrm{Vol}(A))^{\frac{1}{d}} + (\mathrm{Vol}(B))^{\frac{1}{d}}$*. Equivalently, for all* $\lambda \in [0,1]$, $\mathrm{Vol}((1-\lambda)A \oplus \lambda B) \geq \mathrm{Vol}(A)^{1-\lambda} \mathrm{Vol}(B)^{\lambda}$.

Proof. We first show that the two given statements are actually equivalent, as it is not obvious. To see that the former implies the latter, recalling that $\mathrm{Vol}(\alpha A) = \alpha^d \mathrm{Vol}(A)$ for $\alpha \geq 0$,

$$\mathrm{Vol}((1-\lambda)A \oplus \lambda B) \geq \left((1-\lambda)\mathrm{Vol}(A)^{\frac{1}{d}} + \lambda \mathrm{Vol}(B)^{\frac{1}{d}} \right)^d \geq \mathrm{Vol}(A)^{1-\lambda} \mathrm{Vol}(B)^{\lambda}.$$

The last inequality used

$$(1-\lambda)x + \lambda y \geq x^{1-\lambda} y^{\lambda} \text{ for } x, y \geq 0, \tag{6}$$

which follows as log is concave and monotone. Conversely, the latter implies the former by optimizing over $\lambda \in (0,1)$ in the following derivation:

$$\begin{aligned} \mathrm{Vol}(A \oplus B) &= \mathrm{Vol}\left((1-\lambda) \cdot \frac{A}{1-\lambda} \oplus \lambda \cdot \frac{B}{\lambda} \right) \\ &\geq \frac{\mathrm{Vol}(A)^{1-\lambda} \mathrm{Vol}(B)^{\lambda}}{(1-\lambda)^{d(1-\lambda)} \lambda^{d\lambda}} = \left((\mathrm{Vol}(A))^{\frac{1}{d}} + (\mathrm{Vol}(B))^{\frac{1}{d}} \right)^d, \end{aligned}$$

[9] That is, this holds when $A, B \subseteq \mathbb{R}^{d'}$ for $d' > d$, but both are contained in parallel d-dimensional subspaces with nonzero interiors in the subspaces. This extension will be useful later in Lemma 11.

which is omitted to avoid tedium. Now, the latter statement is an application of Theorem 4, with $f \leftarrow \exp(-\chi_A)$, $g \leftarrow \exp(-\chi_B)$, and $h \leftarrow \exp(-\chi_{(1-\lambda)A+\lambda B})$. To verify the condition of Theorem 4 holds, whenever $f(\mathbf{x})^{1-\lambda}g(\mathbf{x}')^{\lambda}$ is nonzero, we have $(1-\lambda)\mathbf{x}+\lambda\mathbf{x}' \in (1-\lambda)A+\lambda B$, so h is nonzero there. Theorem 4 then gives the desired $\mathrm{Vol}((1-\lambda)A \oplus \lambda B) \geq \mathrm{Vol}(A)^{1-\lambda}\mathrm{Vol}(B)^{\lambda}$. □

Interestingly, Theorem 5 does not actually require logconcavity of the indicator functions f, g, h, i.e., it holds for arbitrary compact sets, not just convex ones. We mention that Theorem 5 is not too complicated to establish without using Theorem 4. We give proofs in two simpler settings, and sketch how they may be extended to prove Theorem 5 in general.

Lemma 8. *Theorem 5 is true when $d = 1$.*

Proof. Let $a^+ := \max_{a\in A} a$ and $b^- := \min_{b\in B} b$, which exist by compactness. Then $A \oplus B$ contains $A+\{b^-\}$ and $B+\{a^+\}$, which are disjoint (except at a single point) since $b+a^+ \geq b^-+a$ for any $a \in A$, $b \in B$. This establishes $\mathrm{Vol}(A\oplus B) \geq \mathrm{Vol}(A+\{b^-\})+\mathrm{Vol}(B+\{a^+\}) = \mathrm{Vol}(A)+\mathrm{Vol}(B)$. □

Lemma 9. *Theorem 5 is true when A and B are axis-aligned boxes.*

Proof. Let $A = \prod_{i\in[d]}[0, a_i]$ and $B = \prod_{i\in[d]}[0, b_i]$, for nonnegative $\{a_i, b_i\}_{i\in[d]}$. This is without loss of generality, as shifting boxes does not affect their volume, or the volume of their Minkowski sum. The first characterization in Theorem 5 then follows from the AM-GM inequality:

$$\left(\frac{\mathrm{Vol}(A)}{\mathrm{Vol}(A\oplus B)}\right)^{\frac{1}{d}} + \left(\frac{\mathrm{Vol}(B)}{\mathrm{Vol}(A\oplus B)}\right)^{\frac{1}{d}} = \prod_{i\in[d]}\left(\frac{a_i}{a_i+b_i}\right)^{\frac{1}{d}} + \prod_{i\in[d]}\left(\frac{b_i}{a_i+b_i}\right)^{\frac{1}{d}}$$
$$\leq \left(\frac{1}{d}\sum_{i\in[d]}\frac{a_i}{a_i+b_i}\right) + \left(\frac{1}{d}\sum_{i\in[d]}\frac{b_i}{a_i+b_i}\right) = 1.$$

□

More generally, one can extend Lemma 9 to prove the Brunn-Minkowski inequality for any A, B which are finite collections of disjoint axis-aligned boxes, by induction on the number of boxes. The base case is handled by Lemma 9. Next, suppose Theorem 5 is true when A, B consist of $\leq n$ boxes in total, and consider the case of $n+1$ boxes. We claim there exist translations of A, B and an axis-aligned hyperplane H (i.e., $H = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}_i \geq t\}$ for some $i \in [d]$, $t \in \mathbb{R}$), such that

$$\frac{\mathrm{Vol}(A\cap H)}{\mathrm{Vol}(B\cap H)} = \frac{\mathrm{Vol}(A\cap H^c)}{\mathrm{Vol}(B\cap H^c)} = \frac{\mathrm{Vol}(A)}{\mathrm{Vol}(B)}, \tag{7}$$

and at least one box in A lies entirely in each of H, H^c. To obtain this construction, pick any two[10] (disjoint) boxes in A and any coordinate axis where they are disjoint, choose t appropriately, and shift B to attain (7) along this choice of H. The inductive hypothesis then establishes the desired

$$\begin{aligned}
\mathrm{Vol}(A\oplus B) &\geq \mathrm{Vol}((A\cap H)\oplus(B\cap H)) + \mathrm{Vol}((A\cap H^c)\oplus(B\cap H^c)) \\
&\geq \left(\mathrm{Vol}(A\cap H)^{\frac{1}{d}} + \mathrm{Vol}(B\cap H)^{\frac{1}{d}}\right)^d + \left(\mathrm{Vol}(A\cap H^c)^{\frac{1}{d}} + \mathrm{Vol}(B\cap H^c)^{\frac{1}{d}}\right)^d \\
&= \mathrm{Vol}(A\cap H)\left(1+\left(\frac{\mathrm{Vol}(B\cap H)}{\mathrm{Vol}(A\cap H)}\right)^{\frac{1}{d}}\right)^d + \mathrm{Vol}(A\cap H^c)\left(1+\left(\frac{\mathrm{Vol}(B\cap H^c)}{\mathrm{Vol}(A\cap H^c)}\right)^{\frac{1}{d}}\right)^d \\
&= \mathrm{Vol}(A)\left(1+\left(\frac{\mathrm{Vol}(B)}{\mathrm{Vol}(A)}\right)^{\frac{1}{d}}\right)^d = \left(\mathrm{Vol}(A)^{\frac{1}{d}} + \mathrm{Vol}(B)^{\frac{1}{d}}\right)^d.
\end{aligned}$$

For general sets A, B in Theorem 5, it suffices to take the limit of a sequence of approximations of A, B by collections of disjoint boxes. Finally, we provide a proof of Theorem 4 for completeness.

[10] The problem is symmetric in A, B, and one of A or B has ≥ 2 boxes if we are not in the base case.

Proof of Theorem 4. We induct on d. For $d = 1$, let $L_f(t) := \{x \in \mathbb{R} \mid f(x) \geq t\}$ for all $t \in \mathbb{R}_{\geq 0}$, and similarly define $L_g(t)$ and $L_h(t)$. We observe $L_h(t) \supseteq (1-\lambda)L_f(t) \oplus \lambda L_g(t)$, as

$$h((1-\lambda)x + \lambda x') \geq f(x)^{1-\lambda} g(x')^{\lambda} \geq t \text{ for } x \in L_f(t),\ x' \in L_g(t).$$

Therefore, Lemma 8 shows $\mathrm{Vol}(L_h(t)) \geq (1-\lambda)\mathrm{Vol}(L_f(t)) + \lambda \mathrm{Vol}(L_g(t))$ for all $t \geq 0$. This proves Theorem 4 when $d = 1$, as Fubini's theorem then implies

$$\begin{aligned}
\int h(x)\mathrm{d}x &= \int \left(\int_0^{\infty} \mathbf{1}_{h(x) \geq t} \mathrm{d}t \right) \mathrm{d}x = \int_0^{\infty} \mathrm{Vol}(L_h(t))\mathrm{d}t \\
&\geq (1-\lambda) \int_0^{\infty} \mathrm{Vol}(L_f(t))\mathrm{d}t + \lambda \int_0^{\infty} \mathrm{Vol}(L_g(t))\mathrm{d}t \\
&= (1-\lambda) \int f(x)\mathrm{d}x + \lambda \int g(x)\mathrm{d}x \geq \left(\int f(x)\mathrm{d}x \right)^{1-\lambda} \left(\int g(x)\mathrm{d}x \right)^{\lambda}.
\end{aligned}$$

The last inequality applied (6). Next, for $d > 1$, define for any $a, b \in \mathbb{R}$, and $c := (1-\lambda)a + \lambda b$,

$$f_a(\mathbf{z}) := f(a, \mathbf{z}),\ g_b(\mathbf{z}) := g(b, \mathbf{z}),\ h_c := h(c, \mathbf{z}),\ \text{for all } \mathbf{z} \in \mathbb{R}^{d-1}.$$

Note f_a, g_b, and h_c satisfy Theorem 4's assumption in dimension $d-1$ for any $a, b \in \mathbb{R}$, as

$$h_c((1-\lambda)\mathbf{z} + \lambda \mathbf{z}') = h((1-\lambda)(a, \mathbf{z}) + \lambda(b, \mathbf{z}')) \geq f_a(\mathbf{z})^{1-\lambda} g_b(\mathbf{z}')^{\lambda}.$$

Therefore, by the inductive hypothesis,

$$H(c) := \int_{\mathbb{R}^{d-1}} h_c(\mathbf{z})\mathrm{d}\mathbf{z} \geq \left(\int_{\mathbb{R}^{d-1}} f_a(\mathbf{z})\mathrm{d}\mathbf{z} \right)^{1-\lambda} \left(\int_{\mathbb{R}^{d-1}} g_b(\mathbf{z})\mathrm{d}\mathbf{z} \right)^{\lambda} =: F(a)^{1-\lambda} G(b)^{\lambda}.$$

Hence, the functions F, G, and H defined above satisfy Theorem 4's assumption in dimension 1, so we have the desired conclusion from $\int_{\mathbb{R}} H(a)\mathrm{d}a = \int_{\mathbb{R}^d} h(\mathbf{x})\mathrm{d}\mathbf{x}$, and our base case establishing

$$\int H(a)\mathrm{d}a \geq \left(\int F(a)\mathrm{d}a \right)^{1-\lambda} \left(\int G(a)\mathrm{d}a \right)^{\lambda}.$$

□

5 Grünbaum's theorem

In this section, we establish Theorem 2, which was used in Section 3. Before giving the proof, we state a few useful convex geometry facts, some of which are consequences of the results of Section 4. First, we bound the volume of a cone sliced by a halfplane through its center of gravity.

Lemma 10. *Theorem 2 is true if $S \in \mathbb{R}^d$ is a cone symmetric about $\mathbf{e}_1$, and $\mathbf{v} = \mathbf{e}_1$.*

Proof. Suppose S has base volume and height equal to 1, so its volume is $\frac{1}{d} \cdot 1 \cdot 1 = \frac{1}{d}$. This is without loss of generality by scale invariance of volume ratios. Then its center of gravity is

$$\left(\frac{1}{\mathrm{Vol}(S)} \int_0^1 x \cdot x^{d-1} \mathrm{d}x \right) \mathbf{e}_1 = \left(\frac{1}{\mathrm{Vol}(S)} \cdot \frac{1}{d+1} \right) \mathbf{e}_1 = \frac{d}{d+1} \mathbf{e}_1.$$

Here we used that the base volume at the slice of the cone with first coordinate x scales as x^{d-1}. Now we can compute $\mathrm{Vol}(S \cap H) = (\frac{d}{d+1})^d \mathrm{Vol}(S) \geq \frac{1}{e}\mathrm{Vol}(S)$. □

Next, we prove two useful reductions, which combined with Lemma 10 complete the proof.

Lemma 11. *Let $S \subseteq \mathbb{R}^d$ be convex with $\bar{\mathbf{x}}_S = \mathbf{0}_d$. Consider the construction of a set T, symmetric about $\mathbf{e}_1$, as follows. For each $t \in \mathbb{R}$, let $S_t := S \cap \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}_1 = t\}$, and let $T_t := T \cap \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}_1 = t\}$ be a $(d-1)$-dimensional ball with the same volume as S_t. Then T is convex and $\bar{\mathbf{x}}_T = \bar{\mathbf{x}}_S$.*

Proof. By rotational symmetry of T, the fact that $\bar{\mathbf{x}}_T = \bar{\mathbf{x}}_S$ can be verified along every coordinate other than the first. Moreover, it is clearly true along $\mathbf{e}_1$, since volumes of slices are identical.

Next, note that $\text{rad}(t) \propto \text{Vol}(T_t)^{\frac{1}{d-1}} = \text{Vol}(S_t)^{\frac{1}{d-1}}$, for a normalizing constant depending on $\text{Vol}(\mathbb{B}(\mathbf{0}_d, 1))$. Moreover, we claim $\text{Vol}(S_t)^{\frac{1}{d-1}}$, and hence $\text{rad}(t)$, are concave as functions of t. To see this, for any S_a, S_b, and S_c with $c = (1-\lambda)a + \lambda b$, we have $S_c \supseteq (1-\lambda)S_a \oplus \lambda S_b$ by convexity of S, so Theorem 5 shows the promised concavity of $\text{rad}(t)$:

$$\text{Vol}(S_c)^{\frac{1}{d-1}} \geq (1-\lambda)\text{Vol}(S_a)^{\frac{1}{d-1}} + \lambda \text{Vol}(S_b)^{\frac{1}{d-1}}.$$

Finally, we claim concavity of rad implies that T is convex. To see this, consider two slices T_t and $T_{t'}$ with radii r and r', such that the slice λ along the e_1 axis between r and r' has radius $\geq (1-\lambda)r + \lambda r'$. By convexity of the Euclidean norm (Lemma 4), any point which is the convex combination of points in T_t and $T_{t'}$ then lies in $T_{(1-\lambda)t+\lambda t'}$, as claimed. □

Lemma 12. *Let $T \subseteq \mathbb{R}^d$ be a convex set such that $T_t := T \cap \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}_1 = t\}$ is a $(d-1)$-dimensional ball for all $t \in \mathbb{R}$. Consider the construction of a cone U, symmetric about $\mathbf{e}_1$, as follows. Let $\bar{t} := [\bar{\mathbf{x}}_T]_1$, and choose $t_0 < \bar{t}$ so the cone with tip t_0 and base $T_{\bar{t}}$ has the same volume as $\{\mathbf{x} \in T \mid \mathbf{x}_1 \leq \bar{t}\}$. Finally extend this cone until it has equal volume to T. Then $[\bar{\mathbf{x}}_U]_1 \leq [\bar{\mathbf{x}}_T]_1$.*

Proof. We claim there does not exist $t \in \mathbb{R}$ such that $\text{Vol}(T \cap H_t) > \text{Vol}(U \cap H_t)$, where $H_t := \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}_1 \leq t\}$. This proves that mass has only shifted to the left (along the $\mathbf{e}_1$ direction), which yields the claim. We proceed by contradiction, assuming t is minimal, splitting into two cases.

Case 1: $t \leq \bar{t}$. Let $r(t)$ and $q(t)$ denote the radii of T_t and U_t respectively, i.e., the intersections of T and U with the slice $\{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}_1 = t\}$. Further, $r(t) \geq q(t)$, else we could find a smaller t which maintains $\text{Vol}(T \cap H_t) > \text{Vol}(U \cap H_t)$, contradicting minimality. Let $S_{[t,\bar{t}]} := [\mathbf{x} \in \mathbb{R}^d \mid t \leq \mathbf{x}_1 \leq \bar{t}]$. We claim that $\text{Vol}(T \cap S_{[t,\bar{t}]}) \geq \text{Vol}(U \cap S_{[t,\bar{t}]})$, which follows from concavity of the radius functions r and q (see Lemma 11), since $r(t) \geq q(t)$ and $r(\bar{t}) = q(\bar{t})$. This contradicts our construction, as

$$\text{Vol}(T \cap H_{\bar{t}}) = \text{Vol}(T \cap H_t) + \text{Vol}(T \cap S_{[t,\bar{t}]}) > \text{Vol}(U \cap H_t) + \text{Vol}(U \cap S_{[t,\bar{t}]}) = \text{Vol}(U \cap H_{\bar{t}}).$$

Case 2: $t > \bar{t}$. We claim that there exists $s \in (\bar{t}, t)$ such that $r(s) > q(s)$; otherwise, by minimality of t, we would not have $\text{Vol}(T \cap H_t) > \text{Vol}(U \cap H_t)$. Now $\text{Vol}(T \cap H_s) \leq \text{Vol}(U \cap H_s)$ by minimality of t, so the same contradiction as argued before holds:

$$\text{Vol}(T \cap H_s) = \text{Vol}(T \cap H_{\bar{t}}) + \text{Vol}(T \cap S_{[\bar{t},s]}) > \text{Vol}(U \cap H_{\bar{t}}) + \text{Vol}(U \cap S_{[\bar{t},s]}) = \text{Vol}(U \cap H_s).$$

The only inequality used concavity of r and q, $r(\bar{t}) = q(\bar{t})$, and $r(s) > q(s)$. □

Proof of Theorem 2. Without loss of generality by rotation and shift invariance, let $\mathbf{v} = \mathbf{e}_1$ and $\bar{\mathbf{x}}_S = \mathbf{0}_d$. We then apply the transformations in Lemma 11 and 12 to S, first forming T also with center of gravity $\mathbf{0}_d$, and then forming a cone U with center of gravity $\bar{\mathbf{x}}_U$, such that

$$\text{Vol}(U \cap H_{[\bar{\mathbf{x}}_U]_1}) \leq \text{Vol}(U \cap H_0) = \text{Vol}(S \cap H_0), \text{ and } \text{Vol}(U) = \text{Vol}(S).$$

Finally, Lemma 10 establishes $\text{Vol}(U \cap H_{[\bar{\mathbf{x}}_U]_1}) \geq \frac{1}{e}\text{Vol}(U)$ as claimed. □

Source material

Portions of this lecture are based on reference material in [Roc70, Gar02, Vem11, Bub15, Sid23], as well as the author's own experience working in the field.

References

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[BV04] Dimitris Bertsimas and Santosh S. Vempala. Solving convex programs by random walks. *J. ACM*, 51(4):540–556, 2004.

[CLRS22] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Fourth Edition*. The MIT Press, 2022.

[Gar02] Richard J. Gardner. The brunn-minkowski inequality. *Bulletin of the American Mathematical Society*, 39(3):355–405, 2002.

[Gru60] B. Grunbaum. Partitions of mass-distributions and convex bodies by hyperplanes. *Pacific Journal of Mathematics*, 10(4):1257–1261, 1960.

[JLSW20] Haotian Jiang, Yin Tat Lee, Zhao Song, and Sam Chiu-wai Wong. An improved cutting plane method for convex optimization, convex-concave games, and its applications. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020*, pages 944–953. ACM, 2020.

[Kha80] Leonid G. Khachiyan. Polynomial algorithms in linear programming. *USSR Computational Mathematics and Computational Physics*, 20(1):53–72, 1980.

[LSV18] Yin Tat Lee, Aaron Sidford, and Santosh S. Vempala. Efficient convex optimization with membership oracles. In *Conference On Learning Theory, COLT 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1292–1294. PMLR, 2018.

[LSW15] Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In Venkatesan Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015*, pages 1049–1065. IEEE Computer Society, 2015.

[Pre73] Andras Prekopa. On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 34:335–343, 1973.

[Roc70] R. Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[Sid23] Aaron Sidford. *Optimization Algorithms*. 2023.

[Vai96] Pravin M. Vaidya. A new algorithm for minimizing convex functions over convex sets. *Math. Program.*, 73:291–341, 1996.

[Vem11] Santosh Vempala. *Algorithmic Convex Geometry*. 2011.