CS-GY 9223 I: Lecture 7
Preconditioning, acceleration, coordinate decent, etc.

NYU Tandon School of Engineering, Prof. Christopher Musco

Conditions:

- **Convexity:** f is a convex function, $\mathcal{S}$ is a convex set.
- **Bounded initial distant:**

$$\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R$$

- **Bounded gradients (Lipschitz function):**

$$\|\nabla f(\mathbf{x})\|_2 \leq G \text{ for all } \mathbf{x} \in \mathcal{S}.$$

Theorem

GD Convergence Bound] (Projected) Gradient Descent returns $\hat{\mathbf{x}}$ *with* $f(\hat{\mathbf{x}}) \leq \min_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x}) + \epsilon$ *after*

$$T = \frac{R^2 G^2}{\epsilon^2} \text{ iterations.}$$

$\mathbf{x}^* = \min_{\mathbf{x}} \sum_{i=1}^{T} f_i(\mathbf{x}^*)$ (the offline optimum)

Conditions:

- $f_1, \ldots, f_T$ are all convex.
- Each is G-Lipschitz: for all $\mathbf{x}$, i, $\|\nabla f_i(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$.

Theorem (OGD Regret Bound)

After T steps, $\left[\sum_{i=1}^{T} f_i(\mathbf{x}^{(i)})\right] - \left[\sum_{i=1}^{T} f_i(\mathbf{x}^*)\right] \leq RG\sqrt{T}$. *I.e. the average regret* $\frac{1}{T}\left[\sum_{i=1}^{T} f_i(\mathbf{x}^{(i)})\right]$ *is* $\leq \epsilon$ *after:*

$$T = \frac{R^2 G^2}{\epsilon^2} \text{ iterations.}$$

Conditions:

- Finite sum structure: $f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$, with $f_1, \ldots, f_n$ all convex.
- Lipschitz functions: for all $\mathbf{x}, j$, $\|\nabla f_j(\mathbf{x})\|_2 \leq \frac{G'}{n}$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$.

Theorem (SGD Regret Bound)

Stochastic Gradient Descent returns $\hat{\mathbf{x}}$ *with* $\mathbb{E}[f(\hat{\mathbf{x}})] \leq \min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x}) + \epsilon$ *after*

$$T = \frac{R^2 G'^2}{\epsilon^2} \text{ iterations.}$$

We always have that $G' > G$, *but iterations are typically cheaper by a factor of* n.

Can our convergence bounds be tightened for certain functions? Can they guide us towards faster algorithms?

Goals:

- Improve ϵ dependence below $1/\epsilon^2$.
    - Ideally $1/\epsilon$ or $\log(1/\epsilon)$.
- Reduce or eliminate dependence on G and R.
- Further take advantage of structure in the data (e.g. repetition in features in addition to data points).

Definition (β-smoothness)

A function f is β smooth if, for all $\mathbf{x}, \mathbf{y}$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta \|\mathbf{x} - \mathbf{y}\|_2$$

After some calculus (see Lem. 3.4 in Bubeck's book), this implies:

$$[f(\mathbf{y}) - f(\mathbf{x})] - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \leq \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

For a scalar valued function f, equivalent to $f''(x) \leq \beta$.

Recall from definition of convexity that:

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$$

So now we have an upper and lower bound.

$$0 \leq [f(\mathbf{y}) - f(\mathbf{x})] - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \leq \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

Previously learning rate/step size η depended on G. Now choose it based on β:

$$\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} - \frac{1}{\beta}\nabla f(\mathbf{x}^{(t)})$$

Progress per step of gradient descent:

$$\left[f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)})\right] - \nabla f(\mathbf{x}^{(t)})^T(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}) \leq \frac{\beta}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\|_2^2$$

$$\left[f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)})\right] + \frac{1}{\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2 \leq \frac{\beta}{2}\|\frac{1}{\beta}\nabla f(\mathbf{x}^{(t)})\|_2^2$$

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t+1)}) \geq \frac{1}{2\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2$$

Theorem (GD convergence for β-smooth functions.)

Let f be a β smooth convex function and assume we have $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$. If we run GD for T steps with $\eta = \frac{1}{\beta}$ we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \frac{2\beta R^2}{T-1}$$

Corollary: If $T = O\left(\frac{\beta R^2}{\epsilon}\right)$ we have $f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \epsilon$.

Definition (α-strongly convex)

A convex function f is α-strongly convex if, for all $\mathbf{x}$, $\mathbf{y}$

$$[f(\mathbf{y}) - f(\mathbf{x})] - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \geq \frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

α is a parameter that will depend on our function.

For a twice-differentiable scalar valued function f, equivalent to $f''(x) \geq \alpha$.

Gradient descent for strongly convex functions:

- Choose number of steps T.
- For $i = 1, \ldots, T$:
  - $\eta = \frac{2}{\alpha \cdot (i+1)}$
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}^{(i)}} f(\mathbf{x}^{(i)})$.

Theorem (GD convergence for α-strongly convex functions.)

Let f be an α-strongly convex function and assume we have that, for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$. If we run GD for T steps (with adaptive step sizes) we have:

$$f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \frac{2G^2}{\alpha(T-1)}$$

Corollary: If $T = O\left(\frac{G^2}{\alpha\epsilon}\right)$ we have $f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon$

What if f is both β-smooth and α-strongly convex?

$$\frac{\alpha}{2}\|x-y\|_2^2 \leq \nabla f(x)^T(x-y) - [f(x)-f(y)] \leq \frac{\beta}{2}\|x-y\|_2^2.$$

$$\frac{\alpha}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \nabla f(\mathbf{x})^T(\mathbf{x}-\mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})] \leq \frac{\beta}{2}\|\mathbf{x}-\mathbf{y}\|_2^2.$$

Theorem (GD for β-smooth, α-strongly convex.)

Let f be a β-smooth and α-strongly convex function. If we run GD for T steps (with step size $\eta = \frac{1}{\beta}$) we have:

$$\|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2 \leq e^{-(T-1)\frac{\alpha}{\beta}}\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2$$

$\kappa = \frac{\beta}{\alpha}$ is called the "condition number" of f.

Is it better if κ is large or small?

Converting to more familiar form: Using that fact the $\nabla f(\mathbf{x}^*) = \mathbf{0}$ along with

$$\frac{\alpha}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \nabla f(\mathbf{x})^T(\mathbf{x}-\mathbf{y}) - [f(\mathbf{x}) - f(\mathbf{y})] \leq \frac{\beta}{2}\|\mathbf{x}-\mathbf{y}\|_2^2,$$

we have:

$$\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2 \leq \frac{2}{\alpha}\left[f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*)\right]$$
$$\|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2 \geq \frac{2}{\beta}\left[f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*)\right]$$

Corollary (GD for β-smooth, α-strongly convex.)

Let f be a β-smooth and α-strongly convex function. If we run GD for T steps (with step size $\eta = \frac{1}{\beta}$) we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \frac{\beta}{\alpha} e^{-(T-1)\frac{\alpha}{\beta}} \cdot \left[f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*) \right]$$

Corollary: If $T = O\left(\frac{\beta}{\alpha}\log(\beta/\alpha\epsilon)\right) = O(\kappa \log(\kappa/\epsilon))$ we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \epsilon \left[f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*) \right]$$

Alternative Corollary: If $T = O\left(\frac{\beta}{\alpha}\log(R\beta/\epsilon)\right)$ we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \epsilon$$

Let f be a twice differentiable function from $\mathbb{R}^d \to \mathbb{R}$. Let the Hessian $\mathbf{H} = \nabla^2 f(\mathbf{x})$ contain all of its second derivatives at a point $\mathbf{x}$. So $\mathbf{H} \in \mathbb{R}^{d \times d}$. We have:

$$\mathbf{H}_{i,j} = \left[\nabla^2 f(\mathbf{x})\right]_{i,j} = \frac{\partial^2 f}{\partial x_i x_j}.$$

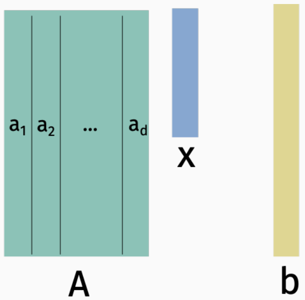
For vector $\mathbf{x}, \mathbf{y}$:

$$\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \approx \left[\nabla^2 f(\mathbf{x})\right](\mathbf{x} - \mathbf{y}).$$

Let f be a twice differentiable function from $\mathbb{R}^d \to \mathbb{R}$. Let the **Hessian** $\mathbf{H} = \nabla^2 f(\mathbf{x})$ contain all of its second derivatives at a point $\mathbf{x}$. So $\mathbf{H} \in \mathbb{R}^{d \times d}$. We have:

$$\mathbf{H}_{i,j} = \left[\nabla^2 f(\mathbf{x})\right]_{i,j} = \frac{\partial^2 f}{\partial x_i x_j}.$$

Example: Let $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$. Recall that $\nabla f(\mathbf{x}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$.



Claim: If f is twice differentiable, then it is convex if and only if the matrix $\mathbf{H} = \nabla^2 f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x}$.

Definition (Positive Semidefinite (PSD))

A square, symmetric matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) for any vector $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{y}^T \mathbf{H} \mathbf{y} \geq 0$.

This is a natural notion of "positivity" for symmetric matrices. To denote that $\mathbf{H}$ is PSD we will typically use "Loewner order" notation (`\succeq` in LaTex):

$$\mathbf{H} \succeq 0.$$

We write $\mathbf{B} \succeq \mathbf{A}$ or equivalently $\mathbf{A} \succeq \mathbf{B}$ to denote that $(\mathbf{B} - \mathbf{A})$ is positive semidefinite. This gives a partial ordering on matrices.

Claim: If f is twice differentiable, then it is convex if and only if the matrix $\mathbf{H} = \nabla^2 f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x}$.

Definition (Positive Semidefinite (PSD))

A square, symmetric matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ is positive semidefinite (PSD) for any vector $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{y}^T \mathbf{H} \mathbf{y} \geq 0$.

For the least squares regression loss function:
$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, $\mathbf{H} = \nabla^2 f(\mathbf{x}) = 2\mathbf{A}^T\mathbf{A}$ for all $\mathbf{x}$. Is $\mathbf{H}$ PSD?

If f is β-smooth and α-strongly convex then at any point $\mathbf{x}$, $\mathbf{H} = \nabla^2 f(\mathbf{x})$ satisfies:

$$\alpha \mathbf{I}_{d\times d} \preceq \mathbf{H} \preceq \beta \mathbf{I}_{d\times d},$$

where $\mathbf{I}_{d\times d}$ is a $d \times d$ identity matrix.

This is the natural matrix generalization of the statement for scalar valued functions:

$$\alpha \leq f''(x) \leq \beta.$$

$$\alpha \mathbf{I}_{d\times d} \preceq \mathbf{H} \preceq \beta \mathbf{I}_{d\times d}.$$

Equivalently for any $\mathbf{z}$,

$$\alpha\|\mathbf{z}\|_2^2 \leq \mathbf{z}^T\mathbf{H}\mathbf{z} \leq \beta\|\mathbf{z}\|_2^2.$$

Exercise: Show that for $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$,

$$[f(\mathbf{x}) - f(\mathbf{y})] - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) = (\mathbf{x} - \mathbf{y})^T \left[2\mathbf{A}^T\mathbf{A}\right] (\mathbf{x} - \mathbf{y}).$$

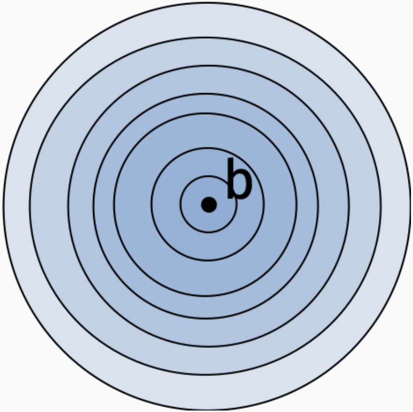
This would imply:

$$\frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \leq [f(\mathbf{x}) - f(\mathbf{y})] - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \leq \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

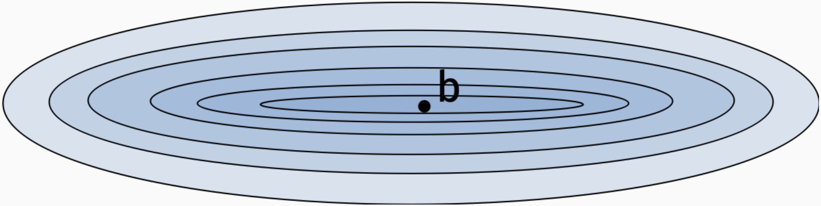
Let $f(\mathbf{x}) = \|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$ where $\mathbf{D}$ is a diagaonl matrix. For now imagine we're in two dimensions: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$.

What are α, β for this problem?

$$\alpha\|\mathbf{z}\|_2^2 \leq \mathbf{z}^T\mathbf{H}\mathbf{z} \leq \beta\|\mathbf{z}\|_2^2$$

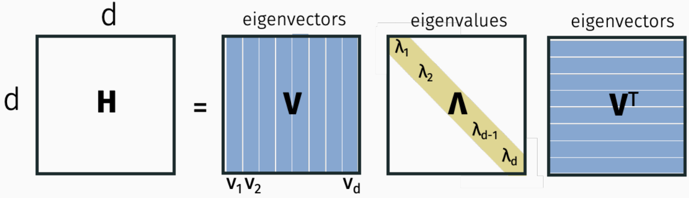


Level sets of $\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$ when $d_1^2 = 1, d_2^2 = 1$.



Level sets of $\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$ when $d_1^2 = \frac{1}{3}, d_2^2 = 2$.

Any symmetric matrix $\mathbf{H}$ has an orthogonal, real valued eigendecomposition.

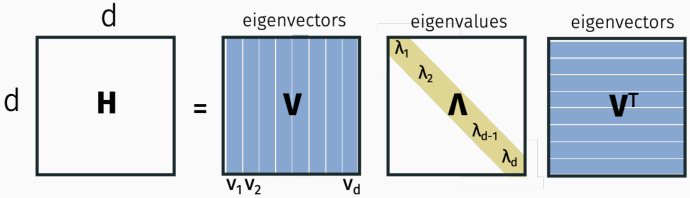


Here $\mathbf{V}$ is square and orthogonal, so $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$. And for each $\mathbf{v}_i$, we have:

$$\mathbf{H}\mathbf{v}_i = \lambda_i \mathbf{v}_i.$$

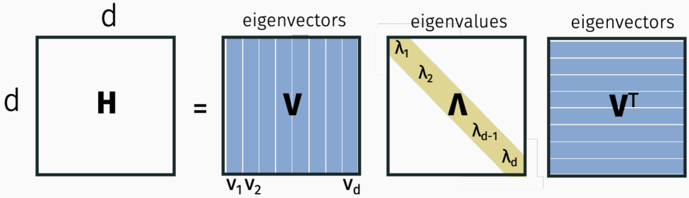
That's what makes $\mathbf{v}_1, \ldots, \mathbf{v}_d$ eigenvectors.

Recall $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.



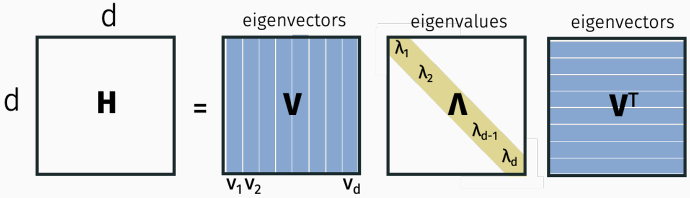
Claim: H $\Leftrightarrow \lambda_1, ..., \lambda_d \geq 0$.

Recall $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.



Claim: $\alpha\mathbf{I} \preceq \mathbf{H} \preceq \beta\mathbf{I} \Leftrightarrow \alpha \leq \lambda_1, ..., \lambda_d \leq \beta$.

Recall $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.



In other words, if we let $\lambda_{\max}(\mathbf{H})$ and $\lambda_{\min}(\mathbf{H})$ be the smallest and largest eigenvalues of $\mathbf{H}$, then for all $\mathbf{z}$ we have:

$$\mathbf{z}^T\mathbf{H}\mathbf{z} \leq \lambda_{\max}(\mathbf{H}) \cdot \|\mathbf{z}\|^2$$
$$\mathbf{z}^T\mathbf{H}\mathbf{z} \geq \lambda_{\min}(\mathbf{H}) \cdot \|\mathbf{z}\|^2$$

If $f(\mathbf{x})$ is β-smooth and α-strongly convex, then for any $\mathbf{x}$ we have the the maximum eigenvalue of $\mathbf{H} = \nabla^2 f(\mathbf{x}) = \beta$ and the minimum eigenvalue of $\mathbf{H} = \nabla^2 f(\mathbf{x}) = \alpha$.

$$\lambda_{\text{max}}(\mathbf{H}) = \beta$$
$$\lambda_{\text{min}}(\mathbf{H}) = \alpha$$

Theorem (GD for β-smooth, α-strongly convex.)

Let f be a β-smooth and α-strongly convex function. If we run GD for T steps (with step size $\eta = \frac{1}{2\beta}$) we have:

$$\|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2 \leq e^{-T/\kappa}\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2$$

Goal: Prove for $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$.

Richardson Iteration view:

$$(\mathbf{x}^{(T+1)} - \mathbf{x}^*) = \left(\mathbf{I} - \frac{1}{\lambda_{\max}}\mathbf{A}^T\mathbf{A}\right)(\mathbf{x}^{(t)} - \mathbf{x}^*)$$

What is the maximum eigenvalue of the symmetric matrix $\left(\mathbf{I} - \frac{1}{\lambda_{\max}}\mathbf{A}^T\mathbf{A}\right)$ in terms of the eigenvalues $\lambda_{\max} = \lambda_1 \geq \ldots \geq \lambda_d = \lambda_{\min}$ of $\mathbf{A}^T\mathbf{A}$?

$$(\mathbf{x}^{(T+1)} - \mathbf{x}^*) = \left(\mathbf{I} - \frac{1}{\lambda_{\max}}\mathbf{A}^T\mathbf{A}\right)^T (\mathbf{x}^{(1)} - \mathbf{x}^*)$$

What is the maximum eigenvalue of the symmetric matrix $\left(\mathbf{I} - \frac{1}{\lambda_{\max}}\mathbf{A}^T\mathbf{A}\right)^T$?

So we have $\|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2 \leq$

We now have a really good understanding of gradient descent.

Number of iterations for ϵ error:

	G-Lipschitz	β-smooth
R bounded start	$O\left(\frac{G^2R^2}{\epsilon^2}\right)$	$O\left(\frac{\beta R^2}{\epsilon}\right)$
α-strong convex	$O\left(\frac{G^2}{\alpha\epsilon}\right)$	$O\left(\frac{\beta}{\alpha}\log(1/\epsilon)\right)$

How do we use this understanding to design faster algorithms?