

Introduction to Convex Optimization for Machine Learning

John Duchi

University of California, Berkeley

Practical Machine Learning, Fall 2009

Outline

What is Optimization

Convex Sets

Convex Functions

Convex Optimization Problems

Lagrange Duality

Optimization Algorithms

Take Home Messages

What is Optimization (and why do we care?)

What is Optimization?

- ▶ Finding the minimizer of a function subject to constraints:

$$\begin{aligned} & \underset{x}{\text{minimize}} && f_0(x) \\ & \text{s.t.} && f_i(x) \leq 0, \quad i = \{1, \dots, k\} \\ & && h_j(x) = 0, \quad j = \{1, \dots, l\} \end{aligned}$$

What is Optimization?

- ▶ Finding the minimizer of a function subject to constraints:

$$\begin{aligned} & \underset{x}{\text{minimize}} && f_0(x) \\ & \text{s.t.} && f_i(x) \leq 0, \quad i = \{1, \dots, k\} \\ & && h_j(x) = 0, \quad j = \{1, \dots, l\} \end{aligned}$$

- ▶ Example: Stock market. “Minimize variance of return subject to getting at least \$50.”

Why do we care?

Optimization is at the heart of many (most practical?) machine learning algorithms.

- ▶ Linear regression:

$$\underset{w}{\text{minimize}} \quad \|Xw - y\|^2$$

- ▶ Classification (logistic regression or SVM):

$$\underset{w}{\text{minimize}} \quad \sum_{i=1}^n \log(1 + \exp(-y_i x_i^T w))$$

$$\text{or } \|w\|^2 + C \sum_{i=1}^n \xi_i \quad \text{s.t.} \quad \xi_i \geq 1 - y_i x_i^T w, \xi_i \geq 0.$$

We still care...

- ▶ Maximum likelihood estimation:

$$\underset{\theta}{\text{maximize}} \quad \sum_{i=1}^n \log p_{\theta}(x_i)$$

- ▶ Collaborative filtering:

$$\underset{w}{\text{minimize}} \quad \sum_{i \prec j} \log (1 + \exp(w^T x_i - w^T x_j))$$

- ▶ k -means:

$$\underset{\mu_1, \dots, \mu_k}{\text{minimize}} \quad J(\mu) = \sum_{j=1}^k \sum_{i \in C_j} \|x_i - \mu_j\|^2$$

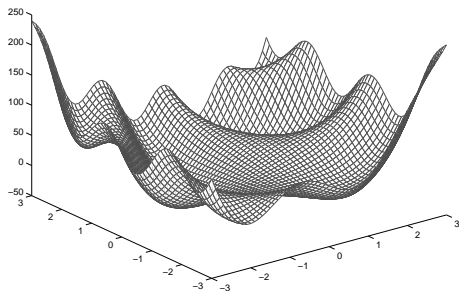
- ▶ And more (graphical models, feature selection, active learning, control)

But generally speaking...

We're screwed.

- ▶ Local (non global) minima of f_0
- ▶ All kinds of constraints (even restricting to continuous functions):

$$h(x) = \sin(2\pi x) = 0$$

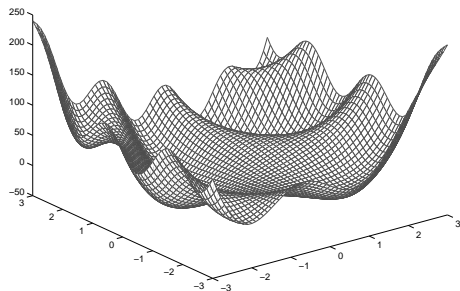


But generally speaking...

We're screwed.

- ▶ Local (non global) minima of f_0
- ▶ All kinds of constraints (even restricting to continuous functions):

$$h(x) = \sin(2\pi x) = 0$$



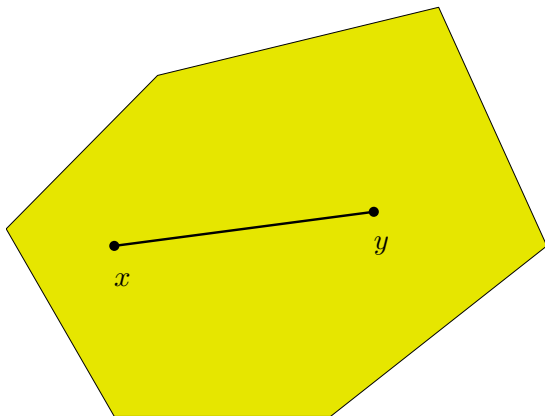
- ▶ Go for convex problems!

Convex Sets

Definition

A set $C \subseteq \mathbb{R}^n$ is *convex* if for $x, y \in C$ and any $\alpha \in [0, 1]$,

$$\alpha x + (1 - \alpha)y \in C.$$



Examples

- ▶ All of \mathbb{R}^n (obvious)

Examples

- ▶ All of \mathbb{R}^n (obvious)
- ▶ Non-negative orthant, \mathbb{R}_+^n : let $x \succeq 0$, $y \succeq 0$, clearly $\alpha x + (1 - \alpha)y \succeq 0$.

Examples

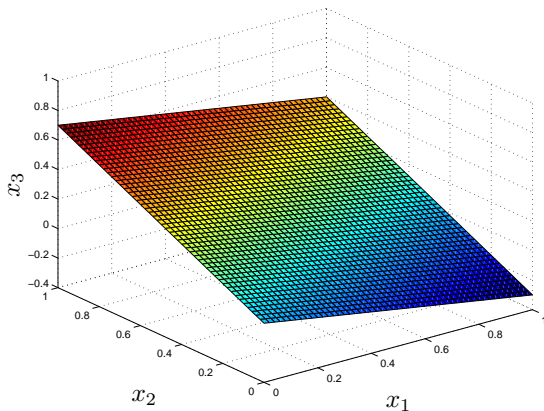
- ▶ All of \mathbb{R}^n (obvious)
- ▶ Non-negative orthant, \mathbb{R}_+^n : let $x \succeq 0$, $y \succeq 0$, clearly $\alpha x + (1 - \alpha)y \succeq 0$.
- ▶ Norm balls: let $\|x\| \leq 1$, $\|y\| \leq 1$, then

$$\|\alpha x + (1 - \alpha)y\| \leq \|\alpha x\| + \|(1 - \alpha)y\| = \alpha \|x\| + (1 - \alpha) \|y\| \leq 1.$$

Examples

- ▶ Affine subspaces: $Ax = b$, $Ay = b$, then

$$A(\alpha x + (1 - \alpha)y) = \alpha Ax + (1 - \alpha)Ay = \alpha b + (1 - \alpha)b = b.$$

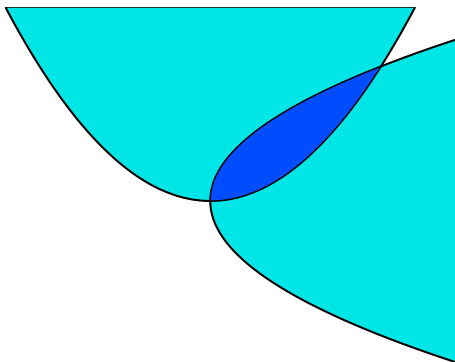


More examples

- ▶ Arbitrary intersections of convex sets: let C_i be convex for $i \in \mathcal{I}$, $C = \bigcap_i C_i$, then

$$x \in C, y \in C \Rightarrow \alpha x + (1 - \alpha)y \in C_i \quad \forall i \in \mathcal{I}$$

so $\alpha x + (1 - \alpha)y \in C$.



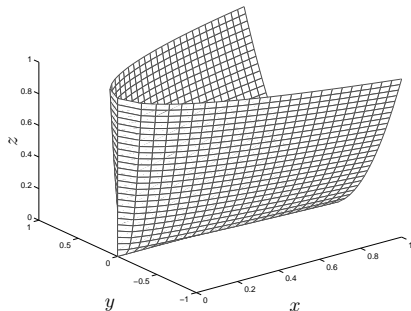
More examples

- ▶ PSD Matrices, a.k.a. the positive semidefinite cone
 $\mathbb{S}_+^n \subset \mathbb{R}^{n \times n}$. $A \in \mathbb{S}_+^n$ means
 $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$. For
 $A, B \in \mathbb{S}_+^n$,

$$\begin{aligned} & x^T (\alpha A + (1 - \alpha) B) x \\ &= \alpha x^T A x + (1 - \alpha) x^T B x \geq 0. \end{aligned}$$

- ▶ On right:

$$\mathbb{S}_+^2 = \left\{ \begin{bmatrix} x & z \\ z & y \end{bmatrix} \succeq 0 \right\} = \{x, y, z : x \geq 0, y \geq 0, xy \geq z^2\}$$

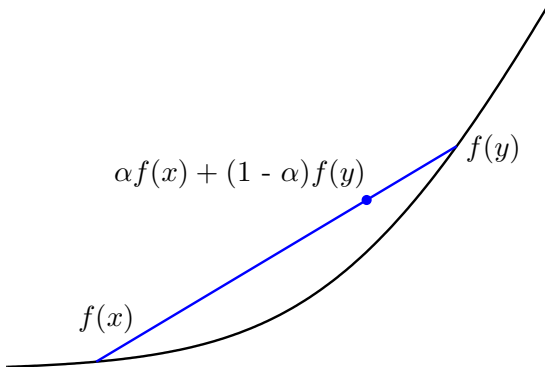


Convex Functions

Definition

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *convex* if for $x, y \in \text{dom } f$ and any $\alpha \in [0, 1]$,

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

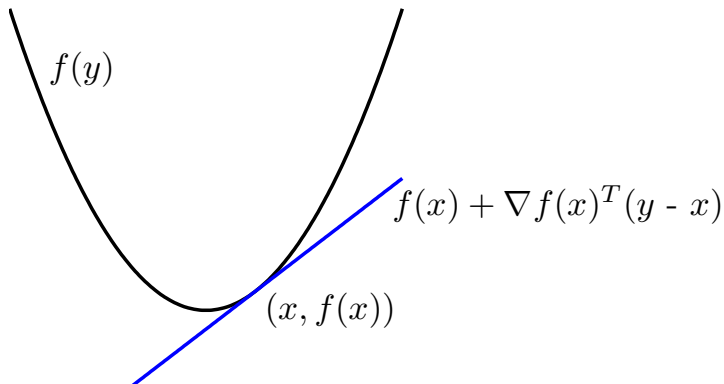


First order convexity conditions

Theorem

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable. Then f is convex if and only if for all $x, y \in \text{dom } f$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$



Actually, more general than that

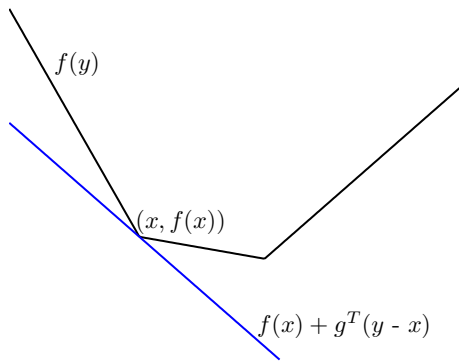
Definition

The *subgradient set*, or subdifferential set, $\partial f(x)$ of f at x is

$$\partial f(x) = \{g : f(y) \geq f(x) + g^T(y - x) \text{ for all } y\}.$$

Theorem

$f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if and only if it has non-empty subdifferential set everywhere.

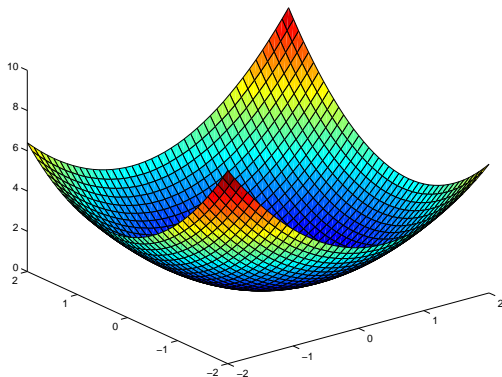


Second order convexity conditions

Theorem

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable. Then f is convex if and only if for all $x \in \text{dom } f$,

$$\nabla^2 f(x) \succeq 0.$$



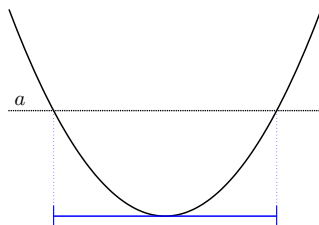
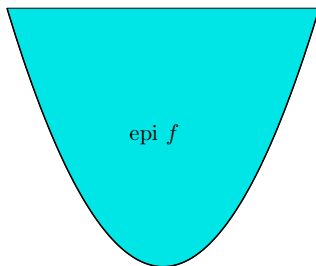
Convex sets and convex functions

Definition

The *epigraph* of a function f is the set of points

$$\text{epi } f = \{(x, t) : f(x) \leq t\}.$$

- ▶ $\text{epi } f$ is convex if and only if f is convex.
- ▶ Sublevel sets, $\{x : f(x) \leq a\}$ are convex for convex f .



Examples

- ▶ Linear/affine functions:

$$f(x) = b^T x + c.$$

Examples

- ▶ Linear/affine functions:

$$f(x) = b^T x + c.$$

- ▶ Quadratic functions:

$$f(x) = \frac{1}{2} x^T A x + b^T x + c$$

for $A \succeq 0$. For regression:

$$\frac{1}{2} \|Xw - y\|^2 = \frac{1}{2} w^T X^T X w - y^T X w + \frac{1}{2} y^T y.$$

More examples

- ▶ Norms (like ℓ_1 or ℓ_2 for regularization):

$$\|\alpha x + (1 - \alpha)y\| \leq \|\alpha x\| + \|(1 - \alpha)y\| = \alpha \|x\| + (1 - \alpha) \|y\| .$$

More examples

- ▶ Norms (like ℓ_1 or ℓ_2 for regularization):

$$\|\alpha x + (1 - \alpha)y\| \leq \|\alpha x\| + \|(1 - \alpha)y\| = \alpha \|x\| + (1 - \alpha) \|y\| .$$

- ▶ Composition with an affine function $f(Ax + b)$:

$$\begin{aligned} f(A(\alpha x + (1 - \alpha)y) + b) &= f(\alpha(Ax + b) + (1 - \alpha)(Ay + b)) \\ &\leq \alpha f(Ax + b) + (1 - \alpha)f(Ay + b) \end{aligned}$$

More examples

- ▶ Norms (like ℓ_1 or ℓ_2 for regularization):

$$\|\alpha x + (1 - \alpha)y\| \leq \|\alpha x\| + \|(1 - \alpha)y\| = \alpha \|x\| + (1 - \alpha) \|y\|.$$

- ▶ Composition with an affine function $f(Ax + b)$:

$$\begin{aligned} f(A(\alpha x + (1 - \alpha)y) + b) &= f(\alpha(Ax + b) + (1 - \alpha)(Ay + b)) \\ &\leq \alpha f(Ax + b) + (1 - \alpha)f(Ay + b) \end{aligned}$$

- ▶ Log-sum-exp (via $\nabla^2 f(x)$ PSD):

$$f(x) = \log \left(\sum_{i=1}^n \exp(x_i) \right)$$

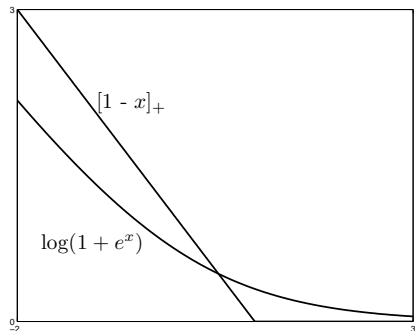
Important examples in Machine Learning

- ▶ SVM loss:

$$f(w) = [1 - y_i x_i^T w]_+$$

- ▶ Binary logistic loss:

$$f(w) = \log(1 + \exp(-y_i x_i^T w))$$



Convex Optimization Problems

Definition

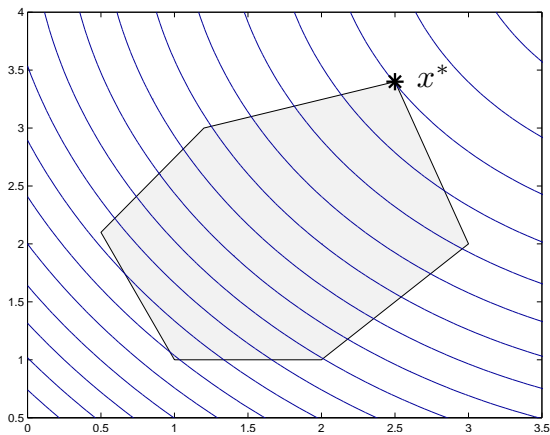
An optimization problem is *convex* if its objective is a convex function, the inequality constraints f_j are convex, and the equality constraints h_j are affine

$$\begin{aligned} & \underset{x}{\text{minimize}} && f_0(x) && \text{(Convex function)} \\ & \text{s.t.} && f_i(x) \leq 0 && \text{(Convex sets)} \\ & && h_j(x) = 0 && \text{(Affine)} \end{aligned}$$

It's nice to be convex

Theorem

If \hat{x} is a local minimizer of a convex optimization problem, it is a global minimizer.



Even more reasons to be convex

Theorem

$\nabla f(x) = 0$ if and only if x is a global minimizer of $f(x)$.

Proof.

- ▶ $\nabla f(x) = 0$. We have

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) = f(x).$$

- ▶ $\nabla f(x) \neq 0$. There is a direction of descent.



LET'S TAKE A BREAK

Lagrange Duality

Goals of Lagrange Duality

- ▶ Get certificate for optimality of a problem
- ▶ Remove constraints
- ▶ Reformulate problem

Constructing the dual

- ▶ Start with optimization problem:

$$\begin{aligned} & \underset{x}{\text{minimize}} && f_0(x) \\ & \text{s.t.} && f_i(x) \leq 0, \quad i = \{1, \dots, k\} \\ & && h_j(x) = 0, \quad j = \{1, \dots, l\} \end{aligned}$$

- ▶ Form *Lagrangian* using Lagrange multipliers $\lambda_i \geq 0$, $\nu_i \in \mathbb{R}$

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^k \lambda_i f_i(x) + \sum_{j=1}^l \nu_j h_j(x)$$

- ▶ Form *dual function*

$$g(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu) = \inf_x \left\{ f_0(x) + \sum_{i=1}^k \lambda_i f_i(x) + \sum_{j=1}^l \nu_j h_j(x) \right\}$$

Remarks

- ▶ Original problem is equivalent to

$$\underset{x}{\text{minimize}} \left[\sup_{\lambda \succeq 0, \nu} \mathcal{L}(x, \lambda, \nu) \right]$$

- ▶ Dual problem is *switching* the min and max:

$$\underset{\lambda \succeq 0, \nu}{\text{maximize}} \left[\inf_x \mathcal{L}(x, \lambda, \nu) \right].$$

One Great Property of Dual

Lemma (Weak Duality)

If $\lambda \succeq 0$, then

$$g(\lambda, \nu) \leq f_0(x^*).$$

Proof.

We have

$$\begin{aligned} g(\lambda, \nu) &= \inf_x \mathcal{L}(x, \lambda, \nu) \leq \mathcal{L}(x^*, \lambda, \nu) \\ &= f_0(x^*) + \sum_{i=1}^k \lambda_i f_i(x^*) + \sum_{j=1}^l \nu_j h_j(x^*) \leq f_0(x^*). \end{aligned}$$


□

The Greatest Property of the Dual

Theorem

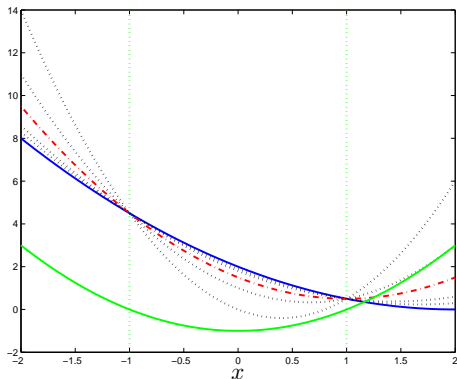
For reasonable¹ convex problems,

$$\sup_{\lambda \succeq 0, \nu} g(\lambda, \nu) = f_0(x^*)$$

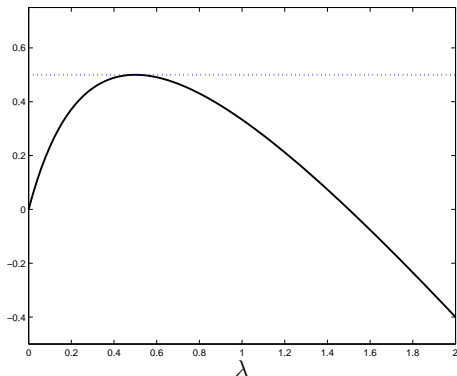
¹There are conditions called constraint qualification for which this is true 

Geometric Look

Minimize $\frac{1}{2}(x - c - 1)^2$ subject to $x^2 \leq c$.



True function (blue), constraint (green), $\mathcal{L}(x, \lambda)$ for different λ (dotted)



Dual function $g(\lambda)$ (black), primal optimal (dotted blue)

Intuition

Can interpret duality as linear approximation.

Intuition

Can interpret duality as linear approximation.

- ▶ $\mathbb{I}_-(a) = \infty$ if $a > 0$, 0 otherwise; $\mathbb{I}_0(a) = \infty$ unless $a = 0$. Rewrite problem as

$$\underset{x}{\text{minimize}} \quad f_0(x) + \sum_{i=1}^k \mathbb{I}_-(f_i(x)) + \sum_{j=1}^l \mathbb{I}_0(h_j(x))$$

Intuition

Can interpret duality as linear approximation.

- ▶ $\mathbb{I}_-(a) = \infty$ if $a > 0$, 0 otherwise; $\mathbb{I}_0(a) = \infty$ unless $a = 0$. Rewrite problem as

$$\underset{x}{\text{minimize}} \quad f_0(x) + \sum_{i=1}^k \mathbb{I}_-(f_i(x)) + \sum_{j=1}^l \mathbb{I}_0(h_j(x))$$

- ▶ Replace $\mathbb{I}(f_i(x))$ with $\lambda_i f_i(x)$; a measure of “displeasure” when $\lambda_i \geq 0$, $f_i(x) > 0$. $\nu_j h_j(x)$ lower bounds $\mathbb{I}_0(h_j(x))$:

$$\underset{x}{\text{minimize}} \quad f_0(x) + \sum_{i=1}^k \lambda_i f_i(x) + \sum_{j=1}^l \nu_j h_j(x)$$

Example: Linearly constrained least squares

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\|^2 \quad \text{s.t.} \quad Bx = d.$$

Form the Lagrangian:

$$\mathcal{L}(x, \nu) = \frac{1}{2} \|Ax - b\|^2 + \nu^T (Bx - d)$$

Take infimum:

$$\nabla_x \mathcal{L}(x, \nu) = A^T Ax - A^T b + B^T \nu \quad \Rightarrow \quad x = (A^T A)^{-1} (A^T b - B^T \nu)$$

Simple unconstrained quadratic problem!

$$\inf_x \mathcal{L}(x, \nu)$$

$$= \frac{1}{2} \|A(A^T A)^{-1} (A^T b - B^T \nu) - b\|^2 + \nu^T B((A^T A)^{-1} A^T b - B^T \nu) - d^T \nu$$

Example: Quadratically constrained least squares

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\|^2 \quad \text{s.t.} \quad \frac{1}{2} \|x\|^2 \leq c.$$

Form the Lagrangian ($\lambda \geq 0$):

$$\mathcal{L}(x, \lambda) = \frac{1}{2} \|Ax - b\|^2 + \frac{1}{2} \lambda (\|x\|^2 - 2c)$$

Take infimum:

$$\nabla_x \mathcal{L}(x, \nu) = A^T Ax - A^T b + \lambda I \quad \Rightarrow \quad x = (A^T A + \lambda I)^{-1} A^T b$$

$$\inf_x \mathcal{L}(x, \lambda) = \frac{1}{2} \|A(A^T A + \lambda I)^{-1} A^T b - b\|^2 + \frac{\lambda}{2} \|(A^T A + \lambda I)^{-1} A^T b\|^2 - \lambda c$$

One variable dual problem!

$$g(\lambda) = -\frac{1}{2} b^T A (A^T A + \lambda I)^{-1} A^T b - \lambda c + \frac{1}{2} \|b\|^2.$$

Uses of the Dual

- ▶ Main use: certificate of optimality (a.k.a. *duality gap*). If we have feasible x and know the dual $g(\lambda, \nu)$, then

$$\begin{aligned} g(\lambda, \nu) \leq f_0(x^*) \leq f_0(x) &\Rightarrow f_0(x^*) - f_0(x) \geq g(\lambda, \nu) - f_0(x) \\ &\Rightarrow f_0(x) - f_0(x^*) \leq f_0(x) - g(\lambda, \nu). \end{aligned}$$

Uses of the Dual

- ▶ Main use: certificate of optimality (a.k.a. *duality gap*). If we have feasible x and know the dual $g(\lambda, \nu)$, then

$$\begin{aligned}g(\lambda, \nu) \leq f_0(x^*) \leq f_0(x) &\Rightarrow f_0(x^*) - f_0(x) \geq g(\lambda, \nu) - f_0(x) \\ &\Rightarrow f_0(x) - f_0(x^*) \leq f_0(x) - g(\lambda, \nu).\end{aligned}$$

- ▶ Also used in more advanced primal-dual algorithms (we won't talk about these).

Optimization Algorithms

Gradient Descent

The simplest algorithm in the world (almost). Goal:

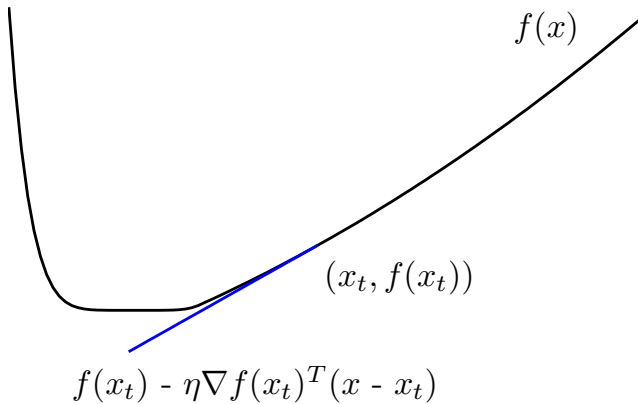
$$\underset{x}{\text{minimize}} \quad f(x)$$

Just iterate

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

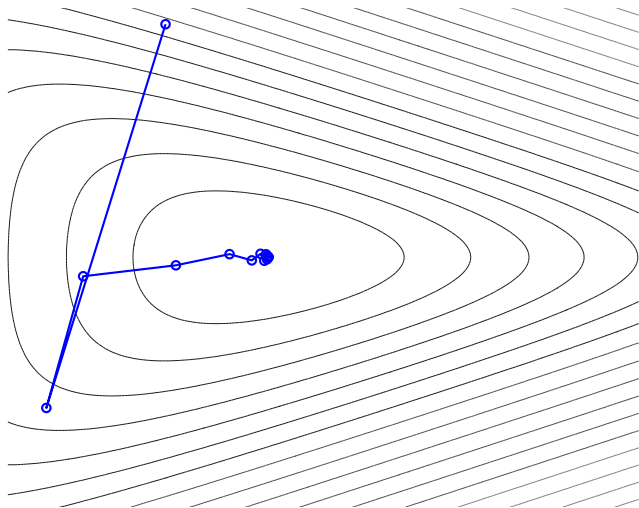
where η_t is stepsize.

Single Step Illustration



Full Gradient Descent

$$f(x) = \log(\exp(x_1 + 3x_2 - .1) + \exp(x_1 - 3x_2 - .1) + \exp(-x_1 - .1))$$



Stepsize Selection

How do I choose a stepsize?

- ▶ Idea 1: exact line search

$$\eta_t = \underset{\eta}{\operatorname{argmin}} f(x - \eta \nabla f(x))$$

Too expensive to be practical.

Stepsize Selection

How do I choose a stepsize?

- ▶ Idea 1: exact line search

$$\eta_t = \underset{\eta}{\operatorname{argmin}} f(x - \eta \nabla f(x))$$

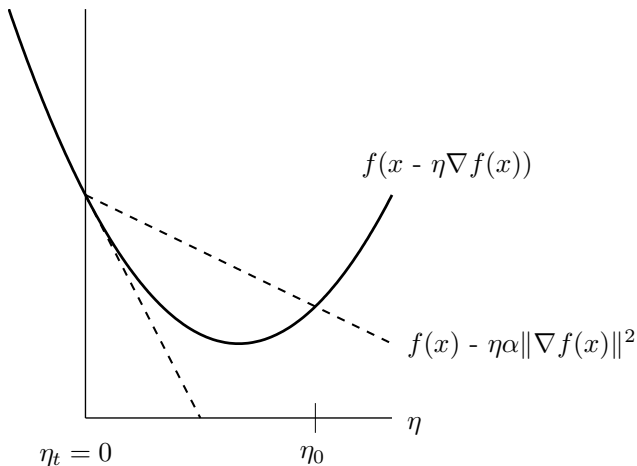
Too expensive to be practical.

- ▶ Idea 2: backtracking (Armijo) line search. Let $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$.
Multiply $\eta = \beta\eta$ until

$$f(x - \eta \nabla f(x)) \leq f(x) - \alpha \eta \|\nabla f(x)\|^2$$

Works well in practice.

Illustration of Armijo/Backtracking Line Search



As a function of stepsize η . Clearly a region where $f(x - \eta \nabla f(x))$ is below line $f(x) - \alpha \eta \|\nabla f(x)\|^2$.

Newton's method

Idea: use a second-order approximation to function.

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x$$

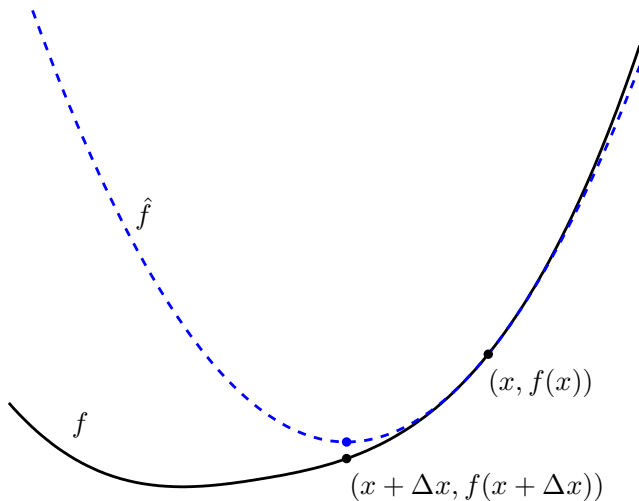
Choose Δx to minimize above:

$$\Delta x = - [\nabla^2 f(x)]^{-1} \nabla f(x)$$

This is descent direction:

$$\nabla f(x)^T \Delta x = -\nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x) < 0.$$

Newton step picture



\hat{f} is 2nd-order approximation, f is true function.

Convergence of gradient descent and Newton's method

- ▶ Strongly convex case: $\nabla^2 f(x) \succeq mI$, then “Linear convergence.” For some $\gamma \in (0, 1)$, $f(x_t) - f(x^*) \leq \gamma^t$, $\gamma < 1$.

$$f(x_t) - f(x^*) \leq \gamma^t \quad \text{or} \quad t \geq \frac{1}{\gamma} \log \frac{1}{\varepsilon} \Rightarrow f(x_t) - f(x^*) \leq \varepsilon.$$

- ▶ Smooth case: $\|\nabla f(x) - \nabla f(y)\| \leq C \|x - y\|$.

$$f(x_t) - f(x^*) \leq \frac{K}{t^2}$$

- ▶ Newton's method often is faster, especially when f has “long valleys”

What about constraints?

- ▶ Linear constraints $Ax = b$ are easy. For example, in Newton method (assume $Ax = b$):

$$\underset{\Delta x}{\text{minimize}} \quad \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x \quad \text{s.t.} \quad A \Delta x = 0.$$

Solution Δx satisfies $A(x + \Delta x) = Ax + A\Delta x = b$.

What about constraints?

- ▶ Linear constraints $Ax = b$ are easy. For example, in Newton method (assume $Ax = b$):

$$\underset{\Delta x}{\text{minimize}} \quad \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x \quad \text{s.t.} \quad A \Delta x = 0.$$

Solution Δx satisfies $A(x + \Delta x) = Ax + A\Delta x = b$.

- ▶ Inequality constraints are a bit tougher...

$$\underset{\Delta x}{\text{minimize}} \quad \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x \quad \text{s.t.} \quad f_i(x + \Delta x) \leq 0$$

just as hard as original.

Logarithmic Barrier Methods

Goal:

$$\underset{x}{\text{minimize}} \quad f_0(x) \quad \text{s.t.} \quad f_i(x) \leq 0, \quad i = \{1, \dots, k\}$$

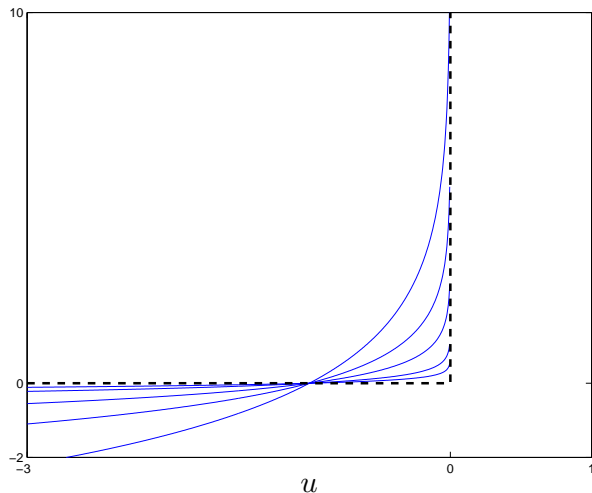
Convert to

$$\underset{x}{\text{minimize}} \quad f_0(x) + \sum_{i=1}^k \mathbb{I}_-(f_i(x))$$

Approximate $\mathbb{I}_-(u) \approx -t \log(-u)$ for small t .

$$\underset{x}{\text{minimize}} \quad f_0(x) - t \sum_{i=1}^k \log(-f_i(x))$$

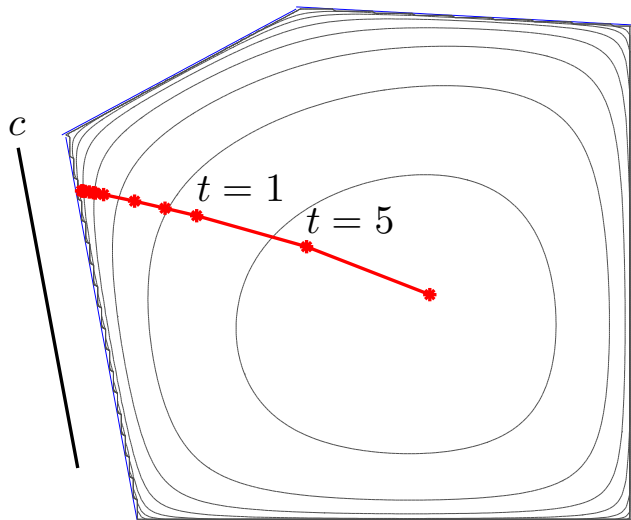
The barrier function



$H_-(u)$ is dotted line, others are $-t \log(-u)$.

Illustration

Minimizing $c^T x$ subject to $Ax \leq b$.



Subgradient Descent

Really, the simplest algorithm in the world. Goal:

$$\underset{x}{\text{minimize}} \quad f(x)$$

Just iterate

$$x_{t+1} = x_t - \eta_t g_t$$

where η_t is a stepsize, $g_t \in \partial f(x_t)$.

Why subgradient descent?

- ▶ Lots of non-differentiable convex functions used in machine learning:

$$f(x) = [1 - a^T x]_+, \quad f(x) = \|x\|_1, \quad f(X) = \sum_{r=1}^k \sigma_r(X)$$

where σ_r is the r th singular value of X .

- ▶ Easy to analyze
- ▶ Do not even need true sub-gradient: just have $\mathbb{E}g_t \in \partial f(x_t)$.

Proof of convergence for subgradient descent

Idea: bound $\|x_{t+1} - x^*\|$ using subgradient inequality. Assume that $\|g_t\| \leq G$.

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \|x_t - \eta g_t - x^*\|^2 \\ &= \|x_t - x^*\|^2 - 2\eta g_t^T (x_t - x^*) + \eta^2 \|g_t\|^2\end{aligned}$$

Proof of convergence for subgradient descent

Idea: bound $\|x_{t+1} - x^*\|$ using subgradient inequality. Assume that $\|g_t\| \leq G$.

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \|x_t - \eta g_t - x^*\|^2 \\ &= \|x_t - x^*\|^2 - 2\eta g_t^T(x_t - x^*) + \eta^2 \|g_t\|^2\end{aligned}$$

Recall that

$$f(x^*) \geq f(x_t) + g_t^T(x^* - x_t) \quad \Rightarrow \quad -g_t^T(x_t - x^*) \leq f(x^*) - f(x_t)$$

so

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 + 2\eta [f(x^*) - f(x_t)] + \eta^2 G^2.$$

Then

$$f(x_t) - f(x^*) \leq \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{2\eta} + \frac{\eta}{2} G^2.$$

Almost done...

Sum from $t = 1$ to T :

$$\begin{aligned}\sum_{t=1}^T f(x_t) - f(x^*) &\leq \frac{1}{2\eta} \sum_{t=1}^T \left[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \right] + \frac{T\eta}{2} G^2 \\ &= \frac{1}{2\eta} \|x_1 - x^*\|^2 - \frac{1}{2\eta} \|x_{T+1} - x^*\|^2 + \frac{T\eta}{2} G^2\end{aligned}$$

Almost done...

Sum from $t = 1$ to T :

$$\begin{aligned} \sum_{t=1}^T f(x_t) - f(x^*) &\leq \frac{1}{2\eta} \sum_{t=1}^T \left[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \right] + \frac{T\eta}{2} G^2 \\ &= \frac{1}{2\eta} \|x_1 - x^*\|^2 - \frac{1}{2\eta} \|x_{T+1} - x^*\|^2 + \frac{T\eta}{2} G^2 \end{aligned}$$

Now let $D = \|x_1 - x^*\|$, and keep track of min along run,

$$f(x_{\text{best}}) - f(x^*) \leq \frac{1}{2\eta T} D^2 + \frac{\eta}{2} G^2.$$

Set $\eta = \frac{D}{G\sqrt{T}}$ and

$$f(x_{\text{best}}) - f(x^*) \leq \frac{DG}{\sqrt{T}}.$$

Extension: projected subgradient descent

Now have a convex constraint set X .

Goal:

$$\underset{x \in X}{\text{minimize}} \quad f(x)$$

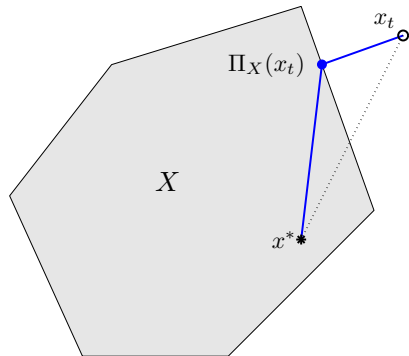
Idea: do subgradient steps, project x_t back into X at every iteration.

$$x_{t+1} = \Pi_X(x_t - \eta g_t)$$

Proof:

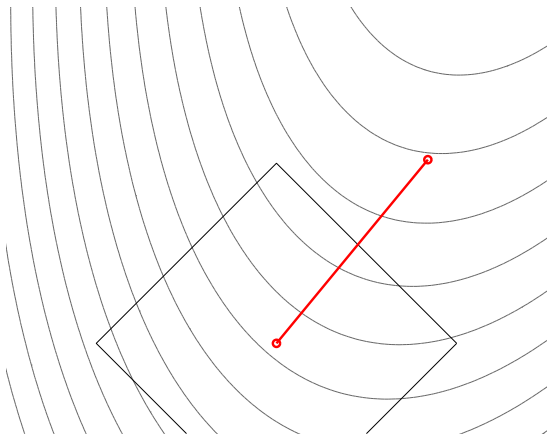
$$\|\Pi_X(x_t) - x^*\| \leq \|x_t - x^*\|$$

if $x^* \in X$.



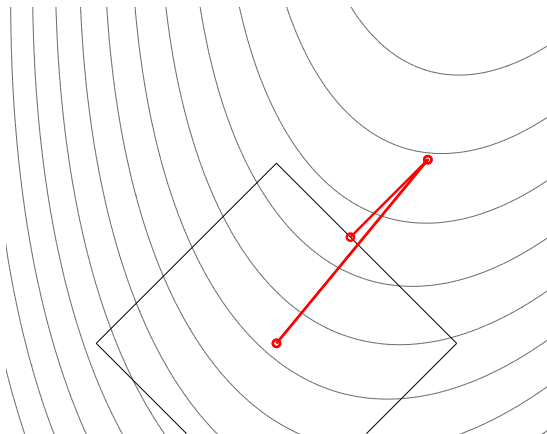
Projected subgradient example

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\| \quad \text{s.t.} \quad \|x\|_1 \leq 1$$



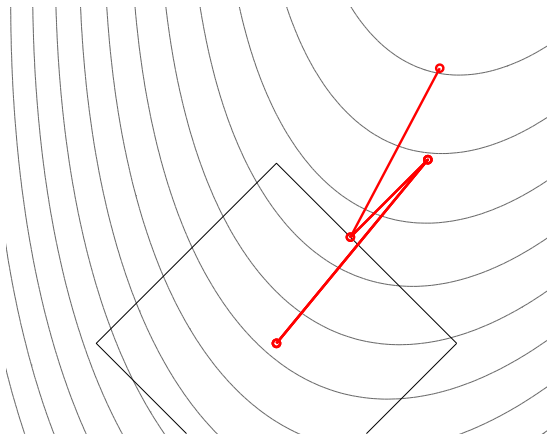
Projected subgradient example

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\| \quad \text{s.t.} \quad \|x\|_1 \leq 1$$



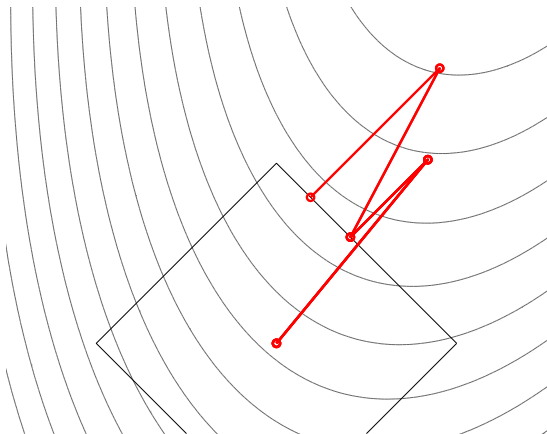
Projected subgradient example

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\| \quad \text{s.t.} \quad \|x\|_1 \leq 1$$



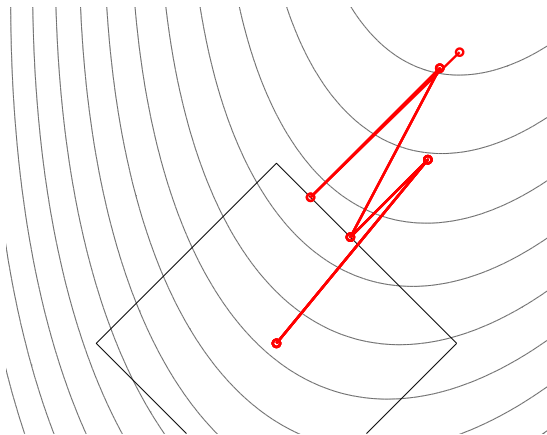
Projected subgradient example

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\| \quad \text{s.t.} \quad \|x\|_1 \leq 1$$



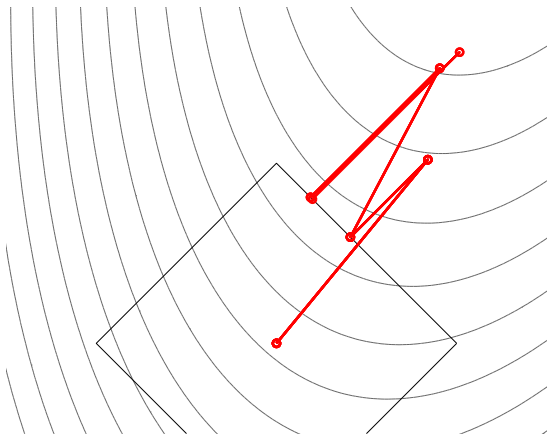
Projected subgradient example

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\| \quad \text{s.t.} \quad \|x\|_1 \leq 1$$



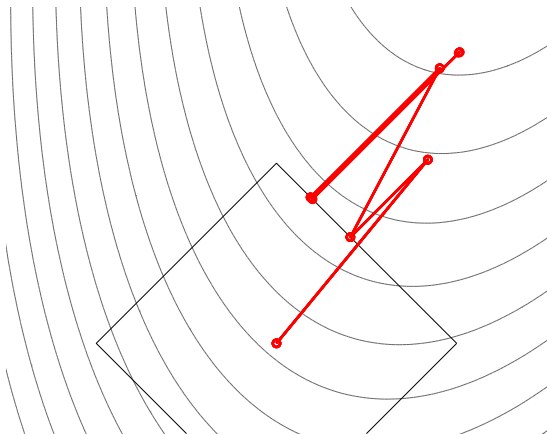
Projected subgradient example

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\| \quad \text{s.t.} \quad \|x\|_1 \leq 1$$



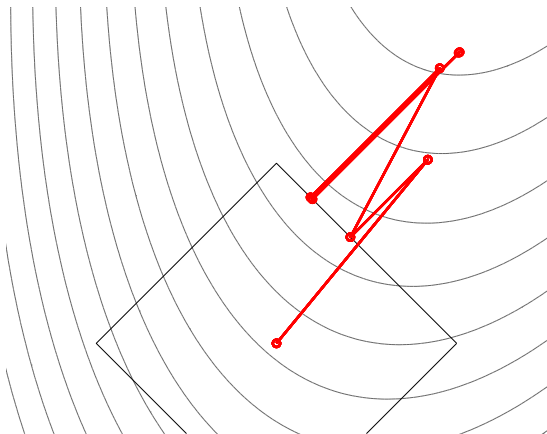
Projected subgradient example

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\| \quad \text{s.t.} \quad \|x\|_1 \leq 1$$



Projected subgradient example

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\| \quad \text{s.t.} \quad \|x\|_1 \leq 1$$



Convergence results for (projected) subgradient methods

- ▶ Any decreasing, non-summable stepsize $\eta_t \rightarrow 0$, $\sum_{t=1}^{\infty} \eta_t = \infty$ gives

$$f(x_{\text{avg}(t)}) - f(x^*) \rightarrow 0.$$

Convergence results for (projected) subgradient methods

- ▶ Any decreasing, non-summable stepsize $\eta_t \rightarrow 0$, $\sum_{t=1}^{\infty} \eta_t = \infty$ gives

$$f(x_{\text{avg}(t)}) - f(x^*) \rightarrow 0.$$

- ▶ Slightly less brain-dead analysis than earlier shows with $\eta_t \propto 1/\sqrt{t}$

$$f(x_{\text{avg}(t)}) - f(x^*) \leq \frac{C}{\sqrt{t}}$$

Convergence results for (projected) subgradient methods

- ▶ Any decreasing, non-summable stepsize $\eta_t \rightarrow 0$, $\sum_{t=1}^{\infty} \eta_t = \infty$ gives

$$f(x_{\text{avg}(t)}) - f(x^*) \rightarrow 0.$$

- ▶ Slightly less brain-dead analysis than earlier shows with $\eta_t \propto 1/\sqrt{t}$

$$f(x_{\text{avg}(t)}) - f(x^*) \leq \frac{C}{\sqrt{t}}$$

- ▶ Same convergence when g_t is random, i.e. $\mathbb{E}g_t \in \partial f(x_t)$. Example:

$$f(w) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n [1 - y_i x_i^T w]_+$$

Just pick random training example.

Recap

- ▶ Defined convex sets and functions
- ▶ Saw why we want optimization problems to be convex (solvable)
- ▶ Sketched some of Lagrange duality
- ▶ First order methods are easy and (often) work well

Take Home Messages

- ▶ Many useful problems formulated as convex optimization problems
- ▶ If it is not convex and not an eigenvalue problem, you are out of luck
- ▶ If it is convex, you are golden