Optimization for Statistics and Machine Learning

Eric Chi

August 12, 2016

What's Duality Good For?

Certify that your algorithm is correct

- duality gap and KKT conditions

Broaden your algorithm design menu

- You could solve the primal problem, but solving the dual problem takes less time to develop code, has better numerical properties, etc.

What's Duality Good For?

Certify that your algorithm is correct

- duality gap and KKT conditions

Broaden your algorithm design menu

- You could solve the primal problem, but solving the dual problem takes less time to develop code, has better numerical properties, etc.

Lagrangian

Standard form optimization problem (not necessarily convex)

$$\begin{aligned} \text{minimize } & f_0(x) \\ \text{subject to } & f_i \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

where $x \in \mathbb{R}^n$, domain $\mathcal{D}$, and optimal value is $p^\star$.

Lagrangian: $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, dom $\mathcal{L} = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$,

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- weighted sum of objective and constraint functions
- $\lambda_i \geq 0$ is Lagrange multiplier associated with $f_i(x) \leq 0$
- ν_i is Lagrange multiplier associated with $h_i(x) = 0$

Relationship between Lagrangian and Primal Problem

$$\psi_P(x) = \sup_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu)$$

- "P" stands for "primal"
- If x violates any of the primal constraints ($f_i(x) > 0$ or $h_i(x) \neq 0$ for any i) then $\psi_P(x) = \infty$.
- If x satisfies the primal constraints then $\psi_P(x) = f_0(x)$

$$\psi_P(x) = \begin{cases} f_0(x) & \text{if } x \text{ is primal feasible} \\ \infty & \text{o.w.} \end{cases}$$

Equivalent unconstrained minimization

$$\inf_x \psi_P(x) = \inf_x \sup_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu)$$

Lagrangian Dual Function

Lagrange Dual Function: $\psi_D : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ - "D" stands for "dual"

$$\psi_D(\lambda, \nu) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu)$$

$$= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

- ψ_D is concave, because it is the point-wise infimum of a collection of affine functions.

Lagrangian Dual Function

$$\psi_D(\lambda,\nu) = \inf_{x\in\mathcal{D}} \mathcal{L}(x,\lambda,\nu)$$
$$= \inf_{x\in\mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

lower bound property: if $\lambda \geq 0$, then $\psi_D(\lambda,\nu) \leq p^\star$
Proof: $\tilde{x}$ feasible $\implies \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \leq 0$

$$\psi_D(\lambda,\nu) \leq \mathcal{L}(\tilde{x},\lambda,\nu) \leq f_0(\tilde{x})$$

Since $\tilde{x}$ is an arbitrary feasible point, $\psi_D(\lambda,\nu) \leq p^\star$
Optimize lower bound:

$$\sup_{\lambda\geq 0,\nu} \psi_D(\lambda,\nu) = d^\star \leq p^\star$$

Least Euclidean norm solution to least squares

$$\text{minimize } b^T b$$

$$\text{subject to } X^T X b = X^T y$$

dual function

- $\mathcal{L}(b, \nu) = b^T b + \nu^T (X^T X b - X^T y)$
- Minimize $\mathcal{L}$ over b by setting gradient to zero

$$\nabla_b \mathcal{L}(b, \nu) = 2b + X^T X \nu = 0 \implies b = -(1/2) X^T X \nu$$

- Plug optimal b into $\mathcal{L}$ to get ψ_D:

$$\psi_D(\nu) = \mathcal{L}(-(1/2) X^T X \nu, \nu) = -\frac{1}{4} \nu^T (X^T X)^2 \nu - y^T X \nu$$

lower bound property says

$$p^\star \geq -\frac{1}{4} \nu^T (X^T X)^2 \nu - y^T X \nu,$$

for all ν.

Standard Form LP

$$\begin{aligned} &\text{minimize } c^T x \\ &\text{subject to } Ax = b, \quad x \geq 0 \end{aligned}$$

dual function

$$\begin{aligned} \mathcal{L}(x, \lambda, \nu) &= c^T x + \nu^T (Ax - b) - \lambda^T x \\ &= -b^T \nu + (c + A^T \nu - \lambda)^T x \end{aligned}$$

- $\mathcal{L}$ is affine in x

$$\psi_{\mathcal{D}}(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu) = \begin{cases} -b^T \nu & A^T \nu - \lambda + c = 0 \\ -\infty & \text{o.w.} \end{cases}$$

- $\psi_{\mathcal{D}}(\lambda, \nu)$ is linear on affine domain

$\{(\lambda, \nu) : A^T \nu - \lambda + c = 0\} \implies \psi_{\mathcal{D}}(\lambda, \nu)$ is concave.

lower bound property: $p^\star \geq -b^T \nu$ if $A^T \nu + c \geq 0$.

Lagrange Dual and Fenchel Conjugates

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } Ax \leq b, \quad Cx = d \end{aligned}$$

dual function

$$\begin{aligned} \psi_D(\lambda, \nu) &= \inf_{x \in \mathbf{dom} f_0} \left[f_0(x) + (A^T\lambda + C^T\nu)^T x - b^T\lambda - d^T\nu \right] \\ &= -f_0^\star(-A^T\lambda - C^T\nu) - b^T\lambda - d^T\nu. \end{aligned}$$

Fenchel conjugate

$$f^\star(y) = \sup_{x \in \mathbf{dom} f} \; y^T x - f(x)$$

Quadratic Program

primal problem (assume $P \in S^n_{++}$)

$$\begin{aligned} &\text{minimize } x^T P x + q^T x \\ &\text{subject to } Ax \leq b, \quad Cx = d \end{aligned}$$

dual function

$$\begin{aligned} \psi_D(\lambda, \nu) &= -\frac{1}{4}\left[(z-q)^T P^{-1}(z-q)\right] - b^T\lambda - d^T\nu \\ z &= -A^T\lambda - C^T\nu \end{aligned}$$

The Dual Problem

Lagrange dual problem

$$\begin{aligned} &\text{maximize } \psi_D(\lambda, \nu) \\ &\text{subject to } \lambda \geq 0 \end{aligned}$$

- finds best possible lower bound on $p^\star$ that the Lagrange dual function can provide
- the dual problem is a convex optimization problem with optimal value $d^\star$
- λ, ν are dual feasible if $\lambda \geq 0, (\lambda, \nu) \in \text{dom}\psi_D$
- often simplified by making implicit constraint $(\lambda, \nu) \in \text{dom}\psi_D$ explicit

example: standard form LP and its dual

$$\begin{aligned} &\text{minimize } c^T x \\ &\text{subject to } Ax = b, \quad x \geq 0 \end{aligned}$$

$$\begin{aligned} &\text{maximize } -b^T \nu \\ &\text{subject to } A^T \nu + c \geq 0 \end{aligned}$$

Weak and Strong Duality

weak duality: $d^\star \leq p^\star$

- always holds (for convex and nonconvex programs)
- can be used to find nontrivial lower bounds for difficult problems

strong duality: $d^\star = p^\star$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications**

duality gap: (x, λ, ν)

$$\psi_P(x) - \psi_D(\lambda, \nu)$$

Slater's Constraint Qualification

strong duality holds for a convex problem

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq 0, \quad Ax = b \end{aligned}$$

if it is strictly feasible

$$\exists x \in \text{int}\mathcal{D} : f_i(x) < 0, \quad i = 1, \ldots, m, \quad Ax = b$$

- also guarantees that the dual optimum is attained (if $p^\star > -\infty$)
- can be sharpened: e.g. can replace int $\mathcal{D}$ with relint $\mathcal{D}$; linear inequalities do not need to hold with strict inequality, ...
- there are many other types of constraint qualifications

Inequality Form LP

primal problem

$$\begin{aligned} &\text{minimize } c^T x \\ &\text{subject to } Ax \leq b \end{aligned}$$

dual function

$$\psi_D(\lambda) = \inf_x \left((c + A^T\lambda)^T x - b^T\lambda \right) = \begin{cases} -b^T\lambda & A^T\lambda + c = 0 \\ -\infty & \text{o.w.} \end{cases}$$

dual problem

$$\begin{aligned} &\text{maximize } -b^T\nu \\ &\text{subject to } A^T\nu + c = 0 \end{aligned}$$

- from Slater's condition: $p^\star = d^\star$ if $A\tilde{x} < b$ for some $\tilde{x}$
- $p^\star = d^\star$ as long as either the primal or dual problem is feasible

Linear Programs and Duality Gap

Four possibilities:

- Primal optimal, dual optimal: $p^\star = d^\star$ (no gap)
- Primal unbounded, dual infeasible: $p^\star = d^\star = -\infty$ (no gap)
- Primal infeasible, dual unbounded: $p^\star = d^\star = \infty$ (no gap)
- Primal infeasible, dual infeasible: $p^\star = +\infty, d^\star = -\infty$ (infinite gap)

Quadratic Program

primal problem (assume $P \in S^n_{++}$)

$$\begin{array}{ll} \text{minimize} & x^T P x \\ \text{subject to} & Ax \leq b \end{array}$$

dual function

$$\psi_D(\lambda) = \inf_x \left(x^T P x + \lambda^T (Ax - b) \right) = -\frac{1}{4} \lambda^T A P^{-1} A^T \lambda - b^T \lambda$$

dual problem

$$\begin{array}{ll} \text{maximize} & -\frac{1}{4} \lambda^T A P^{-1} A^T \lambda - b^T \lambda \\ \text{subject to} & \lambda \geq 0 \end{array}$$

- from Slater's condition: $p^\star = d^\star$ if $A\tilde{x} < b$ for some $\tilde{x}$
- $p^\star = d^\star$ always

Mind the Duality Gap

Theory: Zero duality gap is a certificate of optimality
Thought Experiment:

- Todd gives Munir a primal solution x and claims it is optimal.
- Munir trusts Todd but everyone can make an honest mistake, even Todd.
- How can Munir check Todd's claim?

Solution:

- Todd can provide not only an x but also (λ, ν).
- Munir plugs x into $\psi_P(x)$ and (λ, ν) into $\psi_D(\lambda, \nu)$
- Does $\psi_P(x) = \psi_D(\lambda, \nu)$?

Mind the Duality Gap

Practical Consequence: Debugging your code / checking for correctness

Scenario 1:

- Your iterative algorithm for solving a problem that admits strong duality has been running for 1000 hours. The duality gap hasn't changed since the first minute of running.
- You probably have a bug in your code.
- You want to see the duality gap tending towards zero.

Mind the Duality Gap

Practical Consequence: Debugging your code / checking for correctness

Scenario 2:

- Someone else's iterative algorithm for solving a problem that admits strong duality has been running for 1000 hours. The duality gap hasn't changed since the first minute of running.
- They probably have a bug in their code.
- You want to see the duality gap tending towards zero.

Mind the Duality Gap

Practical Consequence: Debugging your code / checking for correctness

Scenario 3:

- Your collaborator / advisor doesn't trust the results from your estimator. They tell you they suspect your code is buggy.
- You show that the duality gap on some benchmark problems goes to machine precision, verifying that your code is correct.
- This leads to you and your collaborator / advisor to determine a fundamental flaw in your original estimator which inspires a modification that addresses a non-trivial open problem.
- You publish your paper in JASA T&M

Duality for Unconstrained Problems

$$\text{minimize } f_0(Ax + b)$$

- dual function is constant:
  $\psi_D = \inf_x \mathcal{L}(x) = \inf_x f_0(Ax + b) = p^\star$
- we have strong duality, but dual isn't helpful

Key Idea: Introduce new variables and equality constraints

Equivalent equality constrained problem

$$\begin{aligned} &\underset{x,y}{\text{minimize}} && f_0(y) \\ &\text{subject to} && Ax + b - y = 0 \end{aligned}$$

Duality for Unconstrained Problems

Equivalent equality constrained problem

$$\begin{aligned} &\underset{x,y}{\text{minimize}} && f_0(y) \\ &\text{subject to} && Ax + b - y = 0 \end{aligned}$$

Dual function

$$\begin{aligned} \psi_D(\nu) &= \inf_{x,y} \left(f_0(y) - \nu^T y + \nu^T A x + b^T \nu \right) \\ &= \begin{cases} -f_0^\star(\nu) + b^T \nu & A^T \nu = 0 \\ -\infty & \text{o.w.} \end{cases} \end{aligned}$$

Dual problem

$$\begin{aligned} &\underset{x,y}{\text{minimize}} && b^T \nu - f_0^\star(\nu) \\ &\text{subject to} && A^T \nu = 0 \end{aligned}$$

Duality for least squares regression

$$\underset{b}{\text{minimize}} \ \|y - Xb\|_2^2$$

$$\underset{b,z}{\text{minimize}} \ \|z\|_2^2$$

$$\text{subject to } y - Xb = z$$

Fenchel conjugate

$$f^\star(v) = \sup_{u \in \text{dom} f} v^T u - f(u)$$

Fenchel conjugate of quadratic

$$f_0(u) = u^T u \qquad f_0^\star(v) = \frac{1}{4} v^T v$$

Dual Problem

$$\underset{\nu}{\text{maximize}} \ y^T \nu - \frac{1}{4}\|\nu\|_2^2$$

$$\text{subject to } X^T \nu = 0$$

Complementary Slackness

Assume strong duality holds, $x^\star$ is primal optimal, $(\lambda^\star, \nu^\star)$ is dual optimal.

$$\begin{aligned} f_0(x^\star) = \psi_D(\lambda^\star, \nu^\star) &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \right) \\ &\leq f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^p \nu_i^\star h_i(x^\star) \\ &\leq f_0(x^\star) \end{aligned}$$

- $x^\star$ minimizes $\mathcal{L}(x, \lambda^\star, \nu^\star)$
- $\lambda_i^\star f_i(x^\star) = 0$ for $i = 1, \ldots, m$ (complementary slackness condition)

$$\lambda_i^\star > 0 \implies f_i(x^\star) = 0, \quad f_i(x^\star) < 0 \implies \lambda_i^\star = 0$$

Karush-Kuhn-Tucker (KKT) conditions

the following four conditions are called KKT conditions (for a problem with differentiable f_i, h_i):

- **primal feasibility:**
  $f_i(x) \leq 0, i = 1, \ldots, m, h_i(x) = 0, i = 1, \ldots, p$
- **dual feasibility:** $\lambda \geq 0$
- **complementary slackness:** $\lambda_i f_i(x) = 0, i = 1, \ldots, m$
- **stationarity:** gradient of Lagrangian vanishes with respect to x:

$$\nabla_x \mathcal{L}(x, \lambda, \nu) = \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x)$$

- if strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

KKT conditions for convex problems

if $\tilde{x}, \tilde{\lambda}, \tilde{\nu}$ satisfy KKT for a convex problem, then they are optimal:

- from complementary slackness: $f_0(\tilde{x}) = \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$
- from stationarity (and convexity): $\psi_D(\tilde{\lambda}, \tilde{\nu}) = \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$ hence, $f_0(\tilde{x}) = \psi_D(\tilde{\lambda}, \tilde{\nu})$

if Slater's condition is satisfied: x is optimal if and only if there exist λ, ν that satisfy KKT conditions

- recall that Slater implies strong duality, and dual optimum is attained
- generalizes optimality condition $\nabla f_0(x) = 0$ for unconstrained problem

KKT Conditions: Nonnegative Least Squares

$$\min_{b \geq 0} f(b) := \frac{1}{2}\|y - Xb\|_2^2$$

Lagrangian:

$$\mathcal{L}(b, \lambda) = \frac{1}{2}\|y - Xb\|_2^2 - \lambda^T b$$

- **primal feasibilty:**
- **dual feasibility:** $\lambda \geq 0$
- **complementary slackness:**
- **stationarity:**

KKT Conditions: Nonnegative Least Squares

$$\nabla f(b) = X^T(Xb - y)$$

$b^\star$ is a global solution iff

$$\begin{aligned} b^\star &\geq 0 \\ \nabla f(b^\star) &\geq 0 \\ [\nabla f(b^\star)]_j b_j^\star &= 0. \end{aligned}$$

From KKT conditions to a stopping rule

KKT conditions:

$$\begin{aligned} b^\star &\geq 0 \\ \nabla f(b^\star) &\geq 0 \\ [\nabla f(b^\star)]_j b_j^\star &= 0. \end{aligned}$$

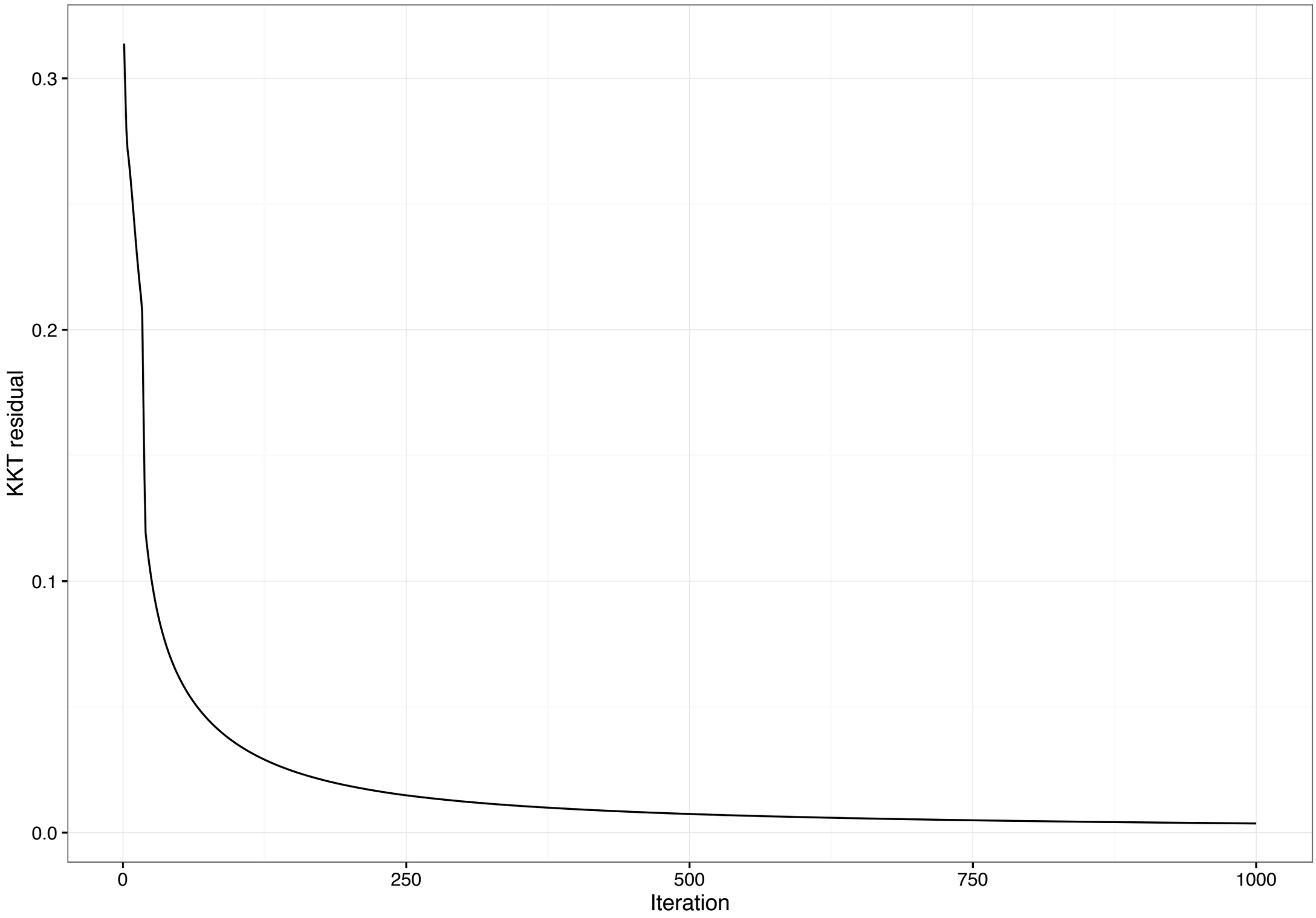
If $b_n \to b^\star$, then

$$\min(b_{nj}, [\nabla f(b_n)]_j) \to 0$$

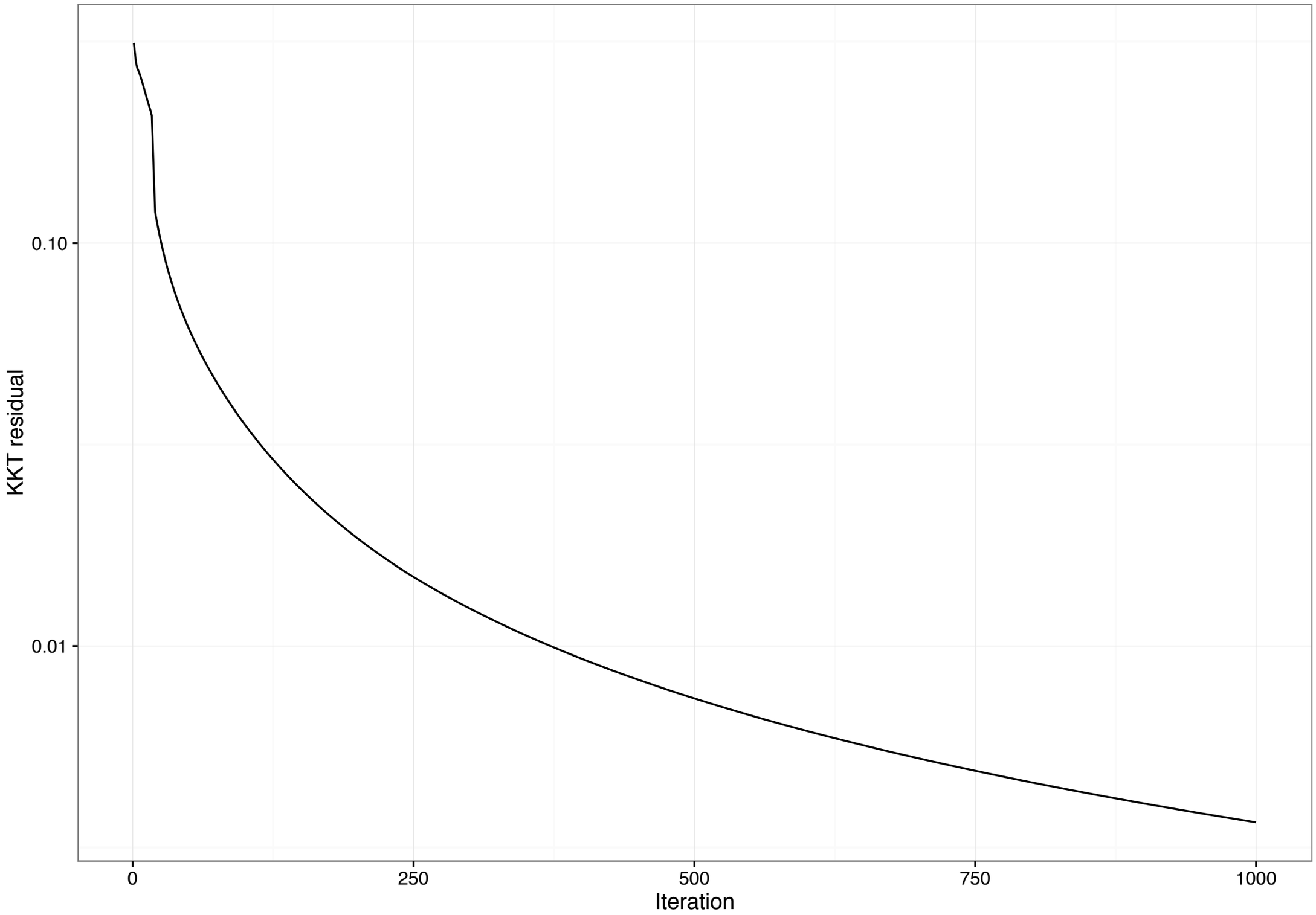
KKT residual:

$$\max_{j=1,\ldots,p} [\min(b_{nj}, [\nabla f(b_n)]_j)] \to 0$$

Track progress with KKT residuals



Track progress with KKT residuals

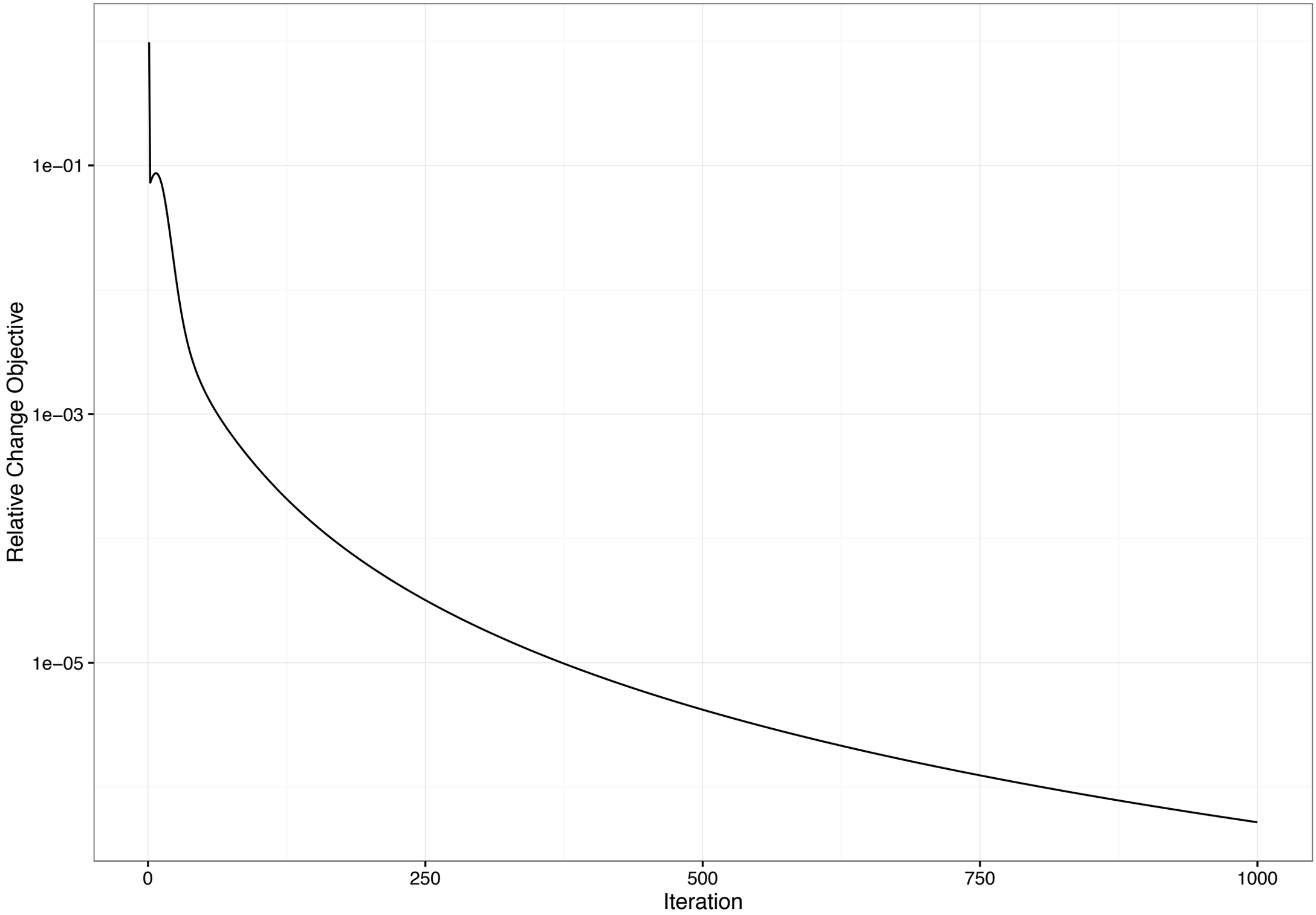


Track progress with relative change in objective

Stop when

$$\frac{f(b) - f(b^+)}{1 + |f(b)|} \leq \text{tolerance}$$

Track progress with relative change in objective

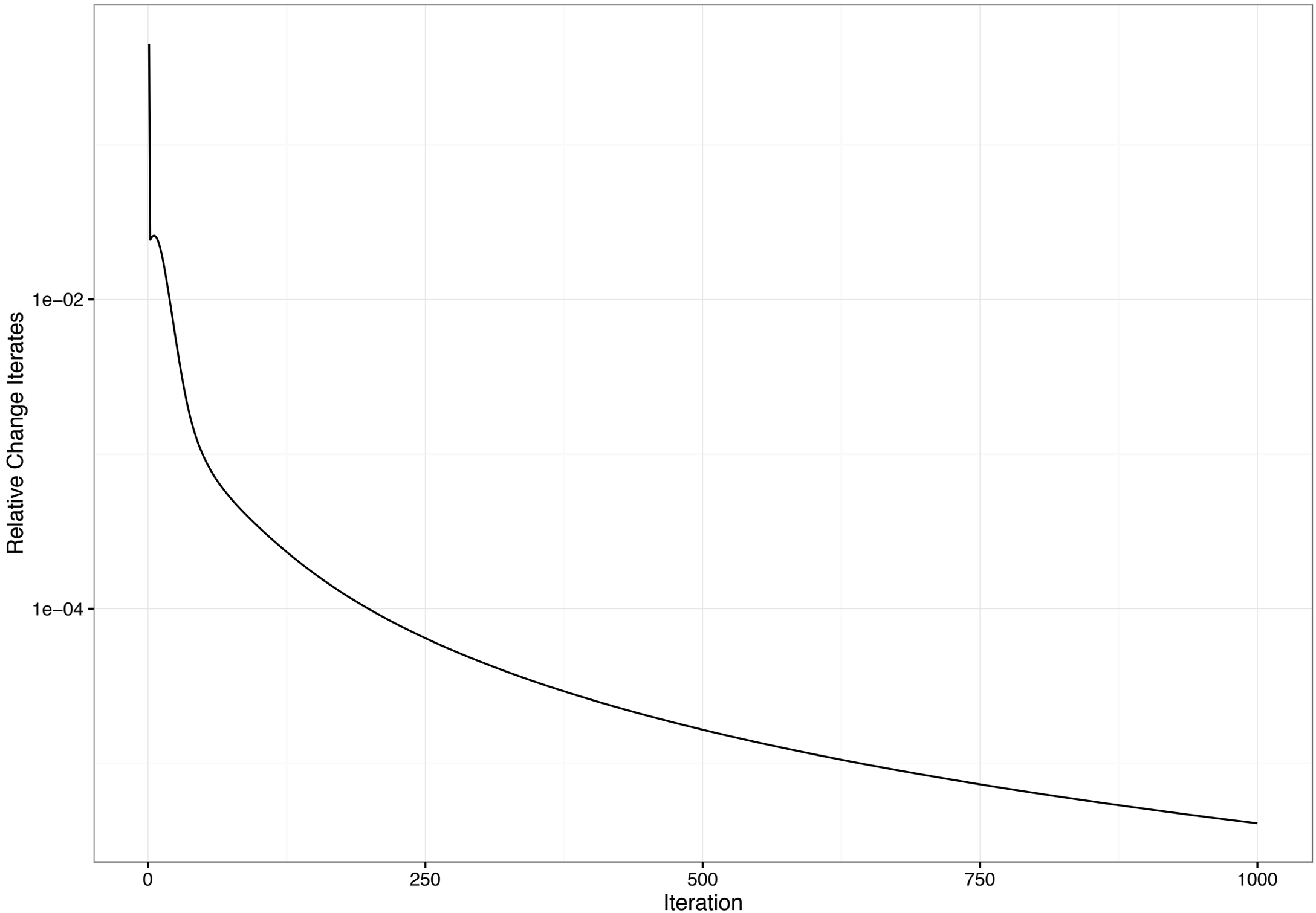


Track progress with relative change in iterates

Stop when

$$\frac{\|b - b^+\|}{1 + \|b\|} \leq \text{tolerance}$$

Track progress with relative change in iterates



Least Absolute Shrinkage and Selection Operator (LASSO)

$$\text{minimize } \|y - Xb\|_2^2 + \gamma \|b\|_1$$

KKT conditions for Lasso

$$\min_{b^+,b^- \geq 0} f(b) := \frac{1}{2}\|y - Xb^+ + Xb^-\|_2^2 + \gamma\langle 1, b^+\rangle + \gamma\langle 1, b^-\rangle$$

Lagrangian:

$$\mathcal{L}(b,\lambda) = \frac{1}{2}\|y - Xb^+ + Xb^-\|_2^2 + \gamma\langle 1, b^+\rangle + \gamma\langle 1, b^-\rangle - \langle \lambda^+, b^+\rangle - \langle \lambda^-, b^-\rangle$$

- **primal feasibilty:** $b^+, b^- \geq 0$
- **dual feasibility:** $\lambda^+, \lambda^- \geq 0$
- **complementary slackness:** $\lambda_j^+ b_j^+ = \lambda_j^- b_j^- = 0$
- **stationarity:**

$$\nabla f(b) + \gamma 1 - \lambda^+ = 0$$
$$-\nabla f(b) + \gamma 1 - \lambda^- = 0$$

KKT conditions for Lasso

- $\|\nabla f(b)\|_\infty \leq \gamma$
- if $b_j > 0$ then $[\nabla f(b)]_j = -\gamma$
- if $b_j < 0$ then $[\nabla f(b)]_j = \gamma$

Q: For what value of γ is $b = 0$ the solution?

Checking convergence:

- Simple / insufficient: $[\|\nabla f(b_n)\|_\infty - \gamma]_+ \to 0$
- More complicated:

$$\max_j \tau(b, j) \to 0$$

where

$$\tau(b,j) = \begin{cases} |[\nabla f(b)]_j + \gamma| & b_j > 0 \\ |[\nabla f(b)]_j - \gamma| & b_j < 0 \\ [|[\nabla f(b)]_j| - \gamma]_+ & b_j = 0 \end{cases}$$

General Principles

- Use KKT conditions (and duality gap) to check your code on small problems for correctness
- In released code, use more lenient stopping criteria (relative change in objective).

Some History on Karush-Kuhn-Tucker

- 1939: William Karush's master's thesis
- 1951: Harold Kuhn and Albert Tucker paper
- 1975: Kuhn sets the record straight by at AMS Symposium on "Nonlinear Programming - A Historical View"

Karush replying to correspondence initiated by Kuhn:

"The thought of publication never occurred to me at the time I wrote the master's thesis. I was a struggling graduate student trying to meet the requirements for going on to my Ph.D. and Graves never brought up the question of publication. I imagine nobody at that time anticipated the future interest in the problem."

What's Duality Good For?

Certify that your algorithm is correct

- duality gap and KKT conditions

→ **Broaden your algorithm design menu**

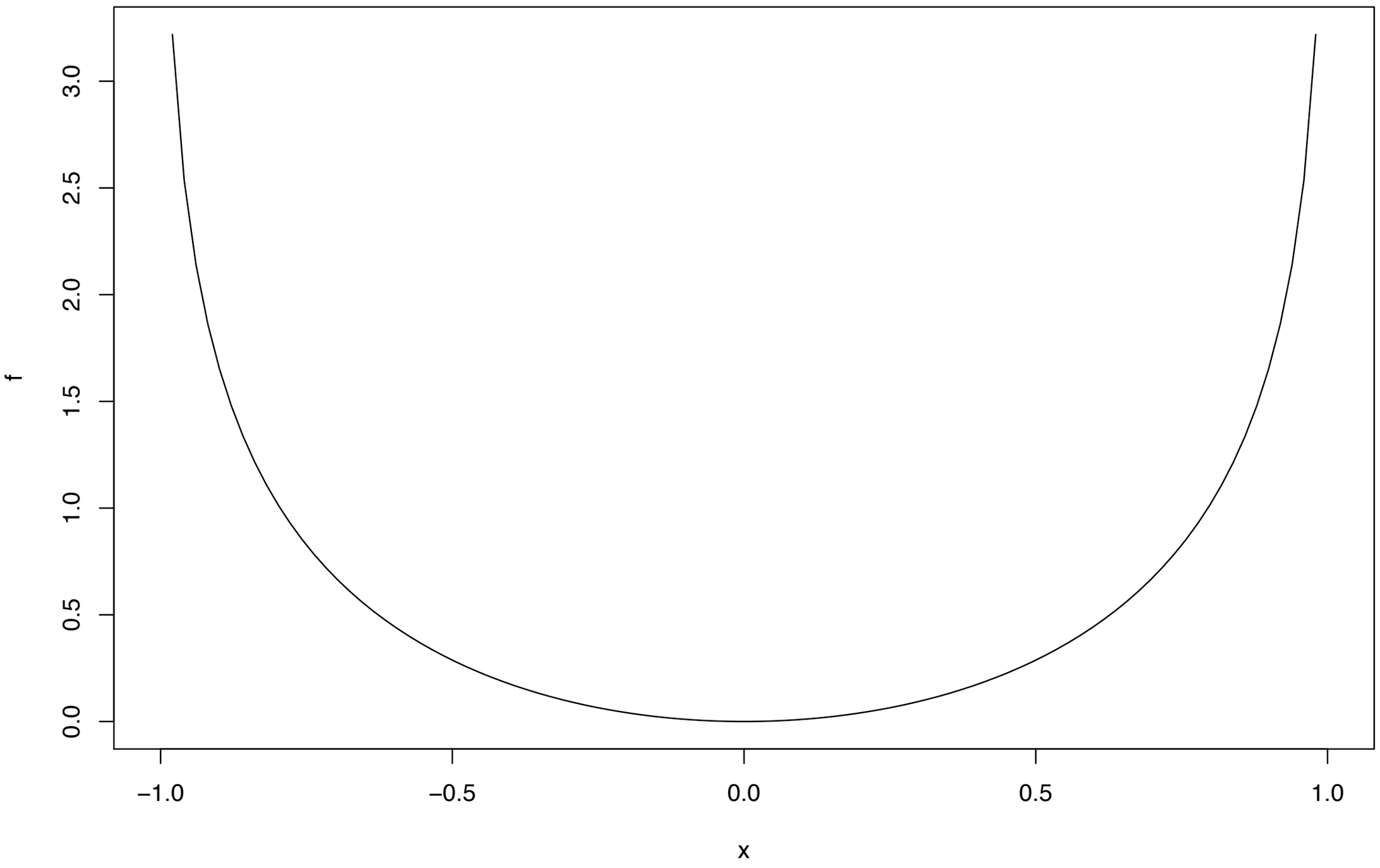
- You could solve the primal problem, but solving the dual problem takes less time to develop code, has better numerical properties, etc.

Closed Convex Functions

Definition: A function f is closed if its epigraph is a closed set

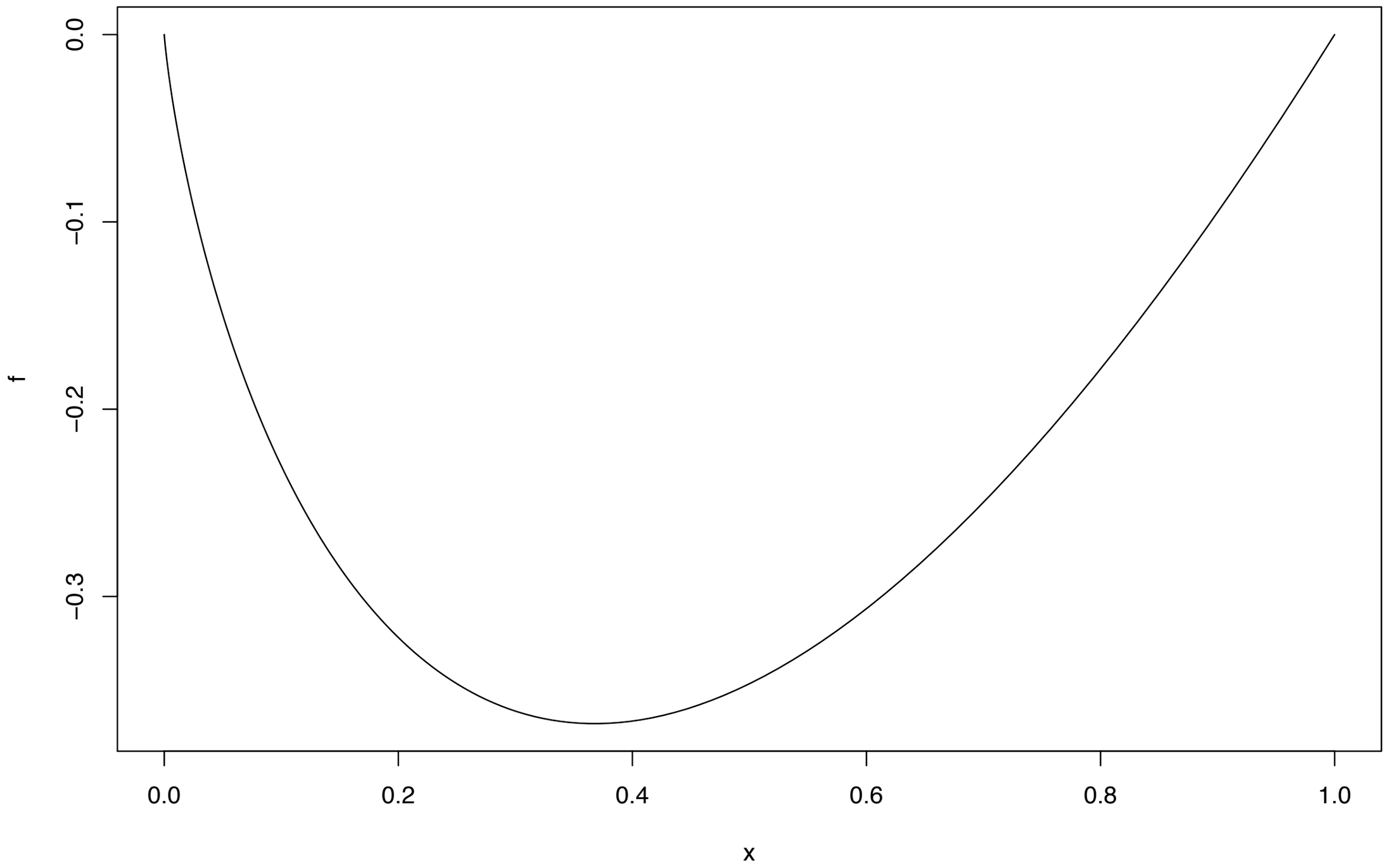
Examples of Closed Convex Functions

$$f(x) = -\log(1 - x^2) \text{ with } \operatorname{dom} f = \{x : |x| < 1\}$$



Examples of Closed Convex Functions

$$f(x) = x \log(x) \text{ with } \operatorname{dom} f = \mathbb{R}_+ \text{ and } f(0) = 0$$



More on Conjugate Functions

Fenchel conjugate of a function f is

$$f^\star(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$$

- $f^\star$ is closed and convex even if f is not
- Fenchel's inequality

$$f(x) + f^\star(y) \geq x^T y, \quad \forall x, y$$

The Second Conjugate

$$f^{\star\star}(x) = \sup_{y \in \mathrm{dom}\, f^\star} (x^T y - f^\star(y))$$

- $f^{\star\star}$ is closed and convex - From Fenchel inequality

$$f^{\star\star}(x) \leq f(x) \quad \forall x$$

- if f is closed and convex, then $f^{\star\star} = f$

Conjugates and Subgradients

if f is closed and convex, then

$$y \in \partial f(x) \iff x^T y = f(x) + f^\star(y) \iff x \in \partial f^\star(y)$$

proof: if $y \in \partial f(x)$, then for all u

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

Therefore, $f^\star(y) = y^T x - f(x)$.

Conjugates and Subgradients

if f is closed and convex, then

$$y \in \partial f(x) \iff x^T y = f(x) + f^\star(y) \iff x \in \partial f^\star(y)$$

proof: if $x^T y = f(x) + f^\star(y)$, then $f^\star(y) = x^T y - f(x)$

$$\implies 0 \in y - \partial f(x)$$
$$\implies y \in \partial f(x)$$

Conjugates and Subgradients

if f is closed and convex, then

$$y \in \partial f(x) \iff x^T y = f(x) + f^\star(y) \iff x \in \partial f^\star(y)$$

proof: if $x \in \partial f^\star(y)$, then for all v

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

Therefore, $f^{\star\star}(x) = x^T y - f^\star(y)$. Since $f^{\star\star} = f$

$$f(x) + f^\star(y) = x^T y$$

Conjugates and Subgradients

if f is closed and convex, then

$$y \in \partial f(x) \iff x^T y = f(x) + f^\star(y) \iff x \in \partial f^\star(y)$$

proof: if $x^T y = f(x) + f^\star(y)$, then $f^{\star\star}(x) = x^T y - f^\star(y)$, since $f^{\star\star} = f$.

$$\implies 0 \in x - \partial f^\star(y)$$

$$\implies x \in \partial f^\star(y)$$

Dual Norm

dual norm of a norm $\|\cdot\|$ is

$$\|y\|_* = \sup_{\|x\| \leq 1} y^T x$$

$\|y\|_*$ is the support function of the unit ball for $\|\cdot\|$

- common pairs of dual vector norms:

$$\|x\|_2, \|y\|_2, \quad \|x\|_1, \|y\|_\infty, \quad \sqrt{x^T Q x}, \sqrt{y^T Q^{-1} y}\,(Q \in S_{++})$$

- common pairs of dual matrix norms

$$\|X\|_{\mathrm{F}}, \|Y\|_{\mathrm{F}}, \quad \|X\|_{\mathrm{op}} = \sigma_{\max}(X) = \|X\|_2, \|Y\|_* = \sum_i \sigma_i(Y)$$

Conjugate of Norm

conjugate of $f = \|x\|$ is the indicator function of the dual unit norm ball

$$f^\star(y) = \sup_x \left(y^T x - \|x\|\right) = \begin{cases} 0 & \|y\|_* \leq 1 \\ +\infty & \text{o.w.} \end{cases}$$

proof:

- if $\|y\|_* \leq 1$, then (by definition of dual norm) $\forall x$

$$\frac{y^T x}{\|x\|} \leq 1$$

$$y^T x \leq \|x\|$$

  and equality holds when $x = 0$; therefore $\sup_x(y^T x - \|x\|) = 0$
- if $\|y\|_* > 1$ then there exists an x with $\|x\| \leq 1, y^T x > 1$; then

$$f^\star(y) \geq y^T(tx) - \|tx\| = t(y^T x - \|x\|) \to \infty \text{ as } t \to \infty$$

Moreau Decomposition

$$x = \text{prox}_h(x) + \text{prox}_{h^\star}(x) \quad \forall x$$

proof:

$$\begin{aligned} u = \text{prox}_h(x) &\iff x - u \in \partial h(x) \\ &\iff u \in \partial h^\star(x - u) \\ &\iff x - u = \text{prox}_{h^\star}(x) \end{aligned}$$

- Generalizes decomposition by orthogonal projection onto subspaces:

  $$x = P_L(x) + P_{L^\perp}(x)$$

  if L is a subspace, $L^\perp$ its orthogonal complement.

(this is the Moreau decomposition with $h = \delta_L, h^\star = \delta_{L^\perp}$)

Basic rules for Proximal Maps

separable sum: $h(x_1, x_2) = h_1(x_1) + h_2(x_2)$

$$\mathrm{prox}_h(x_1, x_2) = (\mathrm{prox}_{h_1}(x_1), \mathrm{prox}_{h_2}(x_2))$$

scaling and translation of argument: $h(x) = f(\lambda x + a)$ with $\lambda \neq 0$

$$\mathrm{prox}_h(x) = \frac{1}{\lambda}(\mathrm{prox}_{\lambda^2 f}(\lambda x + a) - a)$$

conjugate: for $t > 0$

$$\mathrm{prox}_{th^\star}(x) = x - t\,\mathrm{prox}_{h/t}(x/t)$$

for $t = 1$ this is the Moreau decomposition

Support Function

conjugate of support function of closed convex set is indicator function

$$h(x) = S_C(x) = \sup_{y \in C} x^T y, \quad h^\star(y) = \delta_C(y)$$

prox-operator of support function:

$$\begin{aligned} \mathrm{prox}_{th}(x) &= x - t\,\mathrm{prox}_{h^\star/t}(x/t) \\ &= x - tP_C(x/t) \end{aligned}$$

Norms

conjugate of norm is the indicator function of the dual norm ball:

$$h(x) = \|x\|, \quad h^\star(y) = \delta_B(y) \quad (B = \{y : \|y\|_* \leq 1\})$$

prox-operator of norm

$$\begin{aligned}\text{prox}_{th}(x) &= x - t\,\text{prox}_{h^\star/t}(x/t) \\ &= x - tP_B(x/t)\end{aligned}$$

useful for computing $\text{prox}_{t\|\cdot\|}$ when P_B is inexpensive

Soft-Thresholding Operator

- $h(x) = \|x\|_1, B = \{y : \|y\|_\infty \leq 1\}$

Duality and Conjugates

primal problem ($A \in \mathbb{R}^{m \times n}$, f and g convex)

$$\text{minimize } f(x) + g(Ax)$$

Lagrangian (after introducing new variable $y = Ax$)

$$f(x) + g(y) + \langle z, Ax - y \rangle$$

Dual function

$$\inf_x \ (f(x)) + \langle z, Ax \rangle) + \inf_y \ (g(y) - \langle z, y \rangle) = -f^\star(-A^T z) - g^\star(z)$$

Dual problem

$$\text{maximize } \ -f^\star(-A^T z) - g^\star(z)$$

Example: Equality Constraints

g is indicator for $\{b\}$

Primal:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } Ax = b \end{aligned}$$

$$g^\star(z) = \sup_{y \in \{b\}} \langle y, z \rangle = \langle b, z \rangle$$

Dual:

$$\text{maximize } -f^\star(-A^T z) - \langle b, z \rangle$$

Example: Linear inequality constraints

g is indicator for $\{y : y \leq b\}$

Primal:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } Ax \leq b \end{aligned}$$

$$\begin{aligned} g^\star(z) &= \sup_{y \leq b} \langle y, z \rangle \\ &= \sup_{y-b \leq 0} \langle y - b, z \rangle + \langle b, z \rangle \\ &= \langle b, z \rangle + \delta_C(z), \end{aligned}$$

where $C = \{z : z \geq 0\}$

Dual:

$$\begin{aligned} &\text{maximize } -f^\star(-A^T z) - \langle b, z \rangle \\ &\text{subject to } z \geq 0 \end{aligned}$$

Example: Norm regularization

$$g(y) = \|y - b\|$$

Primal:

$$\text{minimize } f(x) + \|Ax - b\|$$

$$g^\star(z) = \langle b, z\rangle + \delta_C(z),$$

where $C = \{z : \|z\|_\dagger \leq 1\}$

Dual:

$$\text{maximize } -f^\star(-A^T z) - \langle b, z\rangle$$

$$\text{subject to } \|z\|_\dagger \leq 1$$

Dual methods

Apply first-order method to the dual problem

$$\text{maximize} \;\; -f^\star(-A^T z) - g^\star(z)$$

Why might the dual be easier for a first-order method?

- Dual problem is unconstrained or has simple constraints
- Dual objective is differentiable or has a simple non-differentiable term
- Decomposition: exploit separable structure

(Sub-)gradients of conjugate function

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is closed and convex with conjugate

$$f^\star(y) = \sup_x \left(\langle y, x\rangle - f(x)\right)$$

subgradient

- $f^\star$ is subdifferentiable on (at least) **int dom** $f^\star$
- maximizers in the definition of $f^\star(y)$ are subgradients at y

$$y \in \partial f(x) \iff \langle y, x\rangle - f(x) = f^\star(y) \iff x \in \partial f^\star(y)$$

gradient: for strictly convex f, maximizer in definition is unique if it exists

$$\nabla f^\star(y) = \arg\max_x \left(\langle y, x\rangle - f(x)\right)$$

if maximum is attained

Conjugate of strongly convex function

Assume f is closed and **strongly** convex, with parameter $\mu > 0$

- $f^\star$ is defined for all y, namely **dom** $f^\star = \mathbb{R}^n$
- $f^\star$ is differentiable everywhere, with gradient

$$\nabla f^\star(y) = \arg\max_x \; (\langle y, x\rangle - f(x))$$

- $\nabla f^\star$ is Lipschitz continuous with constant $1/\mu$

$$\|\nabla f^\star(y) - \nabla f^\star(\tilde{y})\|_2 \leq \frac{1}{\mu}\|y - \tilde{y}\|_2 \quad \forall y, \tilde{y}$$

Composite structure in the dual

Primal:

$$\text{minimize } f(x) + g(Ax)$$

Dual:

$$\text{maximize } -f^\star(-A^T z) - g^\star(z)$$

dual has the right structure for the proximal gradient method if

- prox-operator of g (or $g^\star$) is cheap (closed form or simple algorithm)
- f is strongly convex because then $f^\star(-A^T z)$ has Lipschitz continuous gradient ($L = \|A\|_2^2/\mu$):

$$\|A\nabla f^\star(-A^T u) - A\nabla f^\star(-A^T v)\|_2 \leq \frac{\|A\|_2^2}{\mu}\|u - v\|_2$$

Dual Proximal Gradient Method

$$z^+ = \mathrm{prox}_{tg^\star}\left(z + tA\nabla f^\star(-A^T z)\right)$$

equivalent expression in terms of f:

$$\hat{x} = \arg\min_x \ (f(x) + \langle z, Ax\rangle)$$
$$z^+ = \mathrm{prox}_{tg^\star}\left(z + tA\hat{x}\right)$$

- if f is separable, calculation of $\hat{x}$ decomposes into independent problems
- step size t constant or from backtracking line search
- if $g^\star$ is separable, calculation of z^+ decomposes into independent problems
- can use accelerated first order methods (FISTA, BB)

Alternating minimization interpretation

Recall conjugate rule for prox-operators: for $t > 0$

$$\mathrm{prox}_{tg^\star}(x) = x - t\mathrm{prox}_{g/t}(x/t)$$

Alternative expression for z-update

$$z^+ = z + t(A\hat{x} - \hat{y})$$

where

$$\begin{aligned}
\hat{x} &= \arg\min_x \ (f(x) + \langle z, Ax\rangle) \\
\hat{y} &= \mathrm{prox}_{t^{-1}g}\,(z/t + A\hat{x}) \\
&= \arg\min_y \ \left(g(y) + \langle z, A\hat{x} - y\rangle + \frac{t}{2}\|A\hat{x} - y\|_2^2\right)
\end{aligned}$$

in each iteration, an alternating minimization of:

- Lagrangian $f(x) + g(y) + \langle z, Ax - y\rangle$ over x
- augmented Lagrangian $f(x) + g(y) + \langle z, Ax - y\rangle + \frac{t}{2}\|Ax - y\|_2^2$

Minimization over intersection of convex sets

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in C_1 \cap \cdots \cap C_m \end{aligned}$$

- f strongly convex
- sets C_i are closed, convex, and easy to project onto

$$g(y_1, \ldots, y_m) = \delta_{C_1}(y_1) + \cdots + \delta_{C_m}(y_m)$$

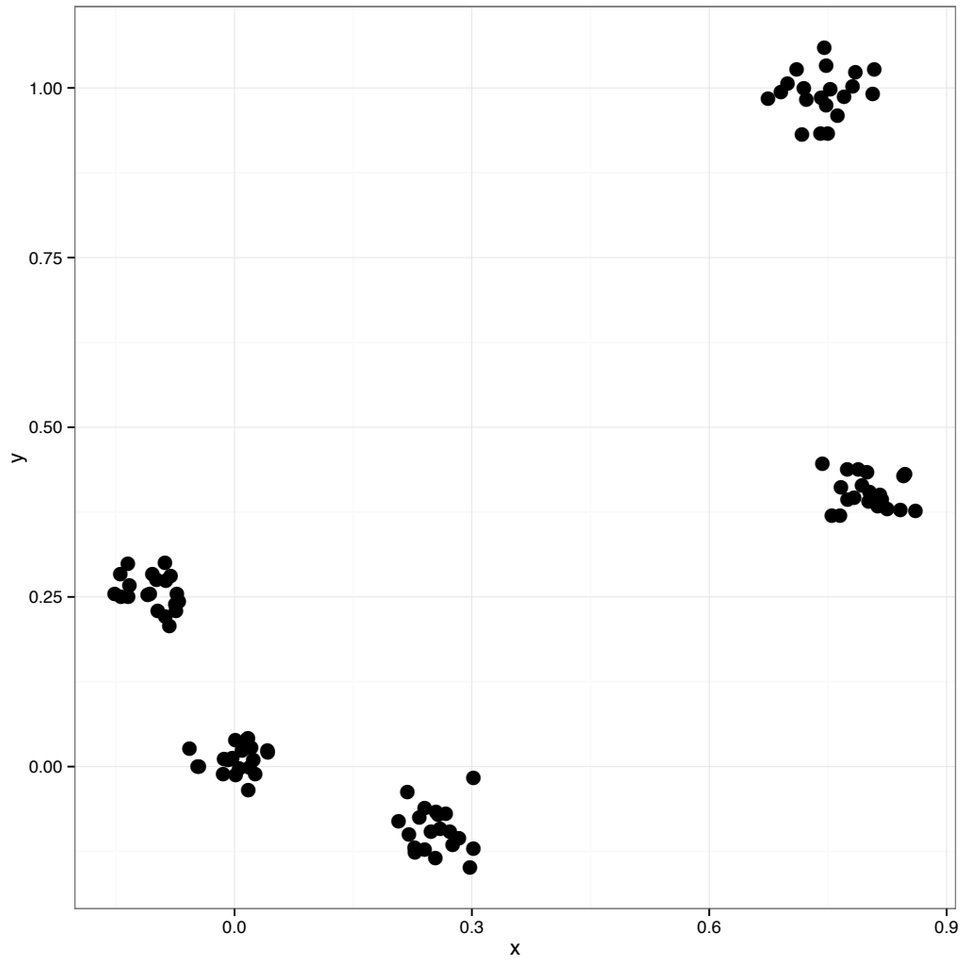
$$A = \begin{pmatrix} I & \cdots & I \end{pmatrix}^T$$

dual proximal gradient update

$$\hat{x} = \arg\min_x \; f(x) + \langle x, \sum_{i=1}^m z_i \rangle$$

$$z_i^+ = z_i + t\hat{x} - tP_{C_i}(z_i/t + \hat{x}), \quad i = 1, \ldots, m$$

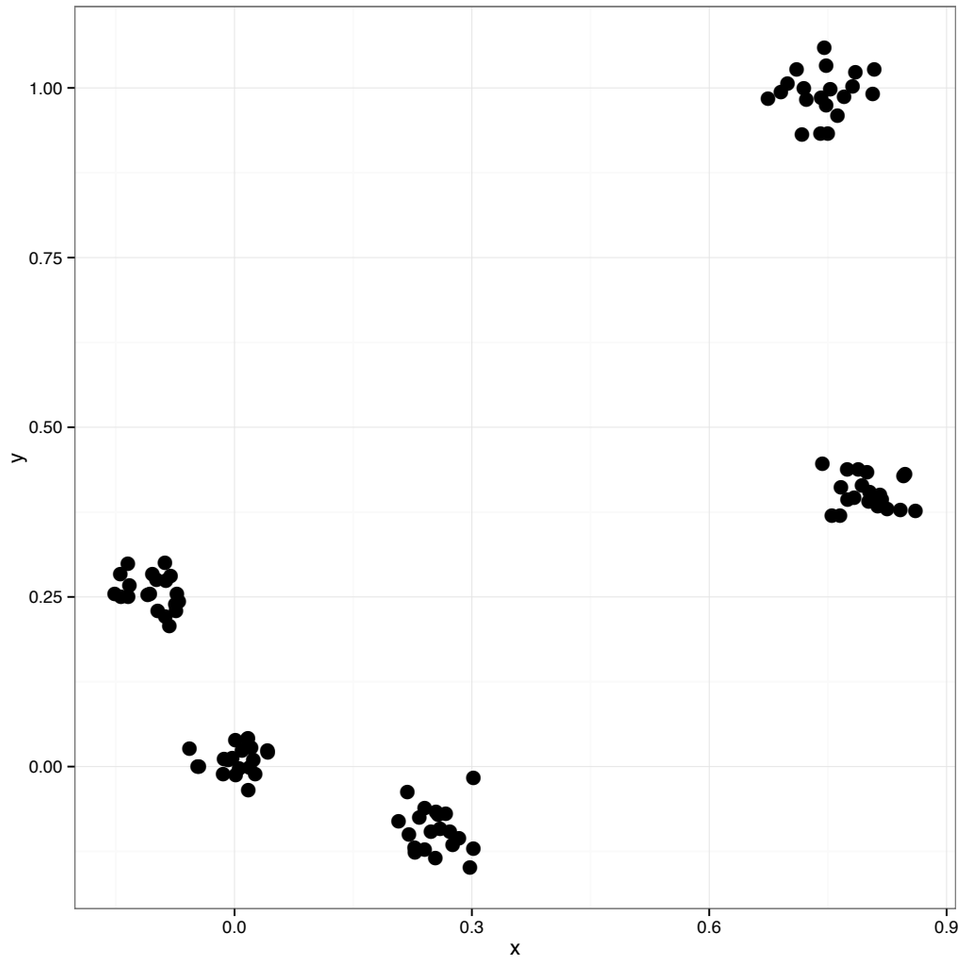
The Clustering Problem



Task:

- Given p points in q dimensions
- $\mathbf{X} \in \mathbb{R}^{q \times p}$
- group similar points together.

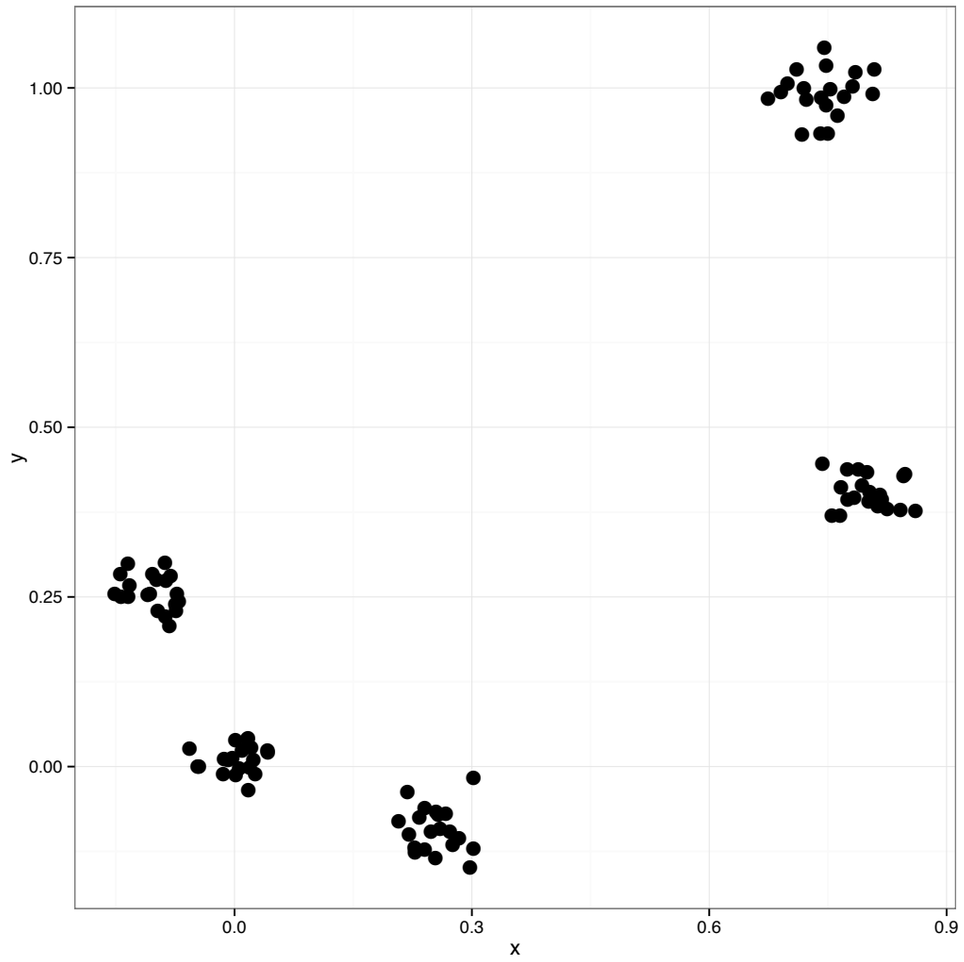
The Clustering Problem



Many approaches:

- k-means, mixture models
- Hierarchical clustering
- Spectral clustering, ...

The Clustering Problem



Computational Issues

- Nonconvex formulations
- Local minimizers
- Instability (initializations, tuning parameters, or data)

Convex Clustering

- Pelckmans et al. 2005, Lindsten et al. 2011, Hocking et al. 2011

$$\underset{\mathbf{u}}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2$$

- Assign a centroid $\mathbf{u}_i$ to each data point $\mathbf{x}_i$.

Convex Clustering

- Pelckmans et al. 2005, Lindsten et al. 2011, Hocking et al. 2011

$$\underset{\mathbf{u}}{\text{minimize}} \; \frac{1}{2} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2$$

- Assign a centroid $\mathbf{u}_i$ to each data point $\mathbf{x}_i$.

Too many degrees of freedom!

Convex Clustering

- Pelckmans et al. 2005, Lindsten et al. 2011, Hocking et al. 2011

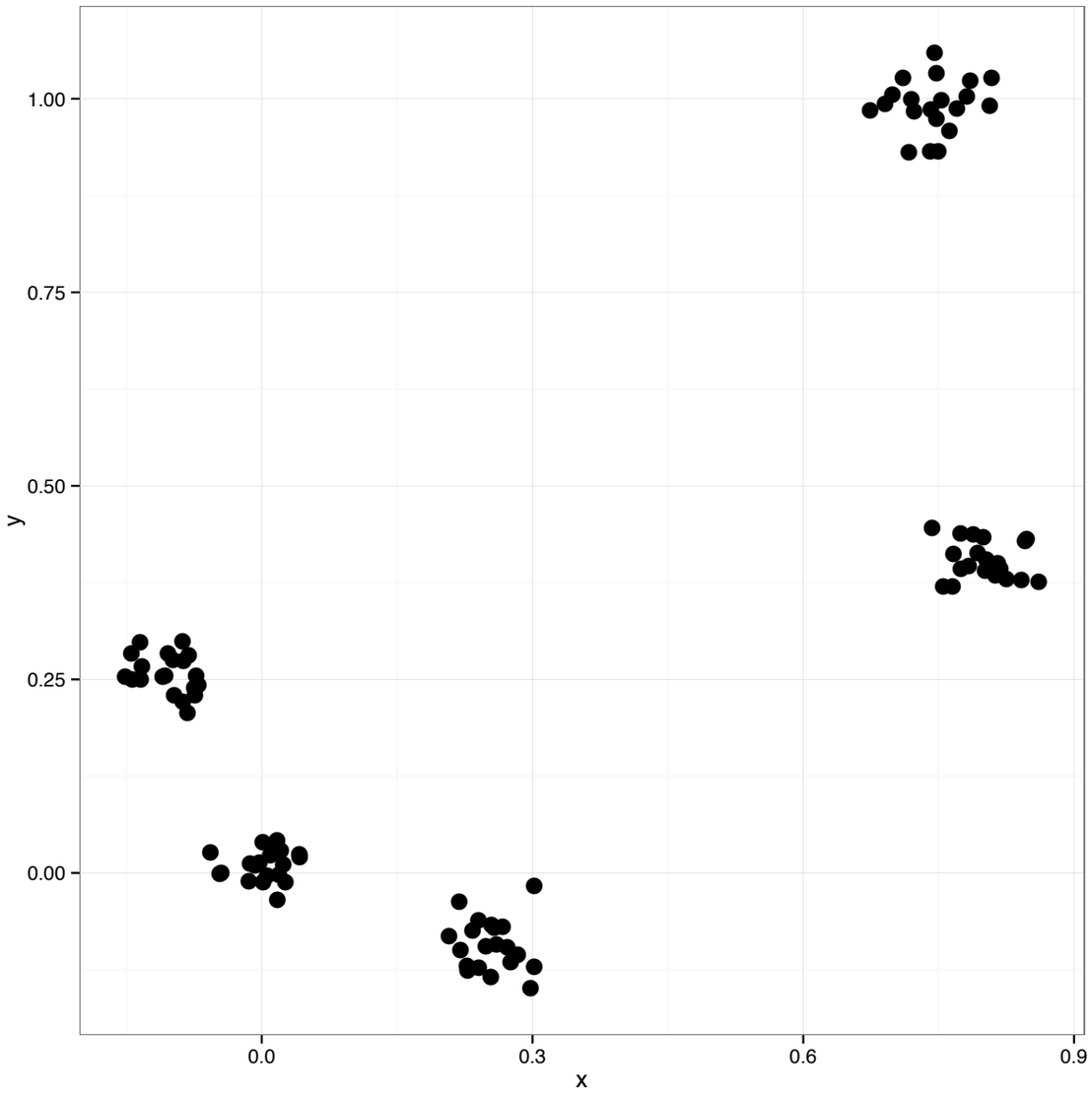
$$\underset{\mathbf{u}}{\text{minimize}} \; \frac{1}{2} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|_2$$

- Assign a centroid $\mathbf{u}_i$ to each data point $\mathbf{x}_i$.
- Convex Fusion Penalty
  - shrinks cluster centroids together
  - sparsity in pairwise differences of centroids

$$\mathbf{u}_i - \mathbf{u}_j = \mathbf{0} \iff \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster}$$

- γ : tunes overall amount of regularization
- w_{ij} : fine tunes pairwise shrinkage
- Generalizes fused lasso

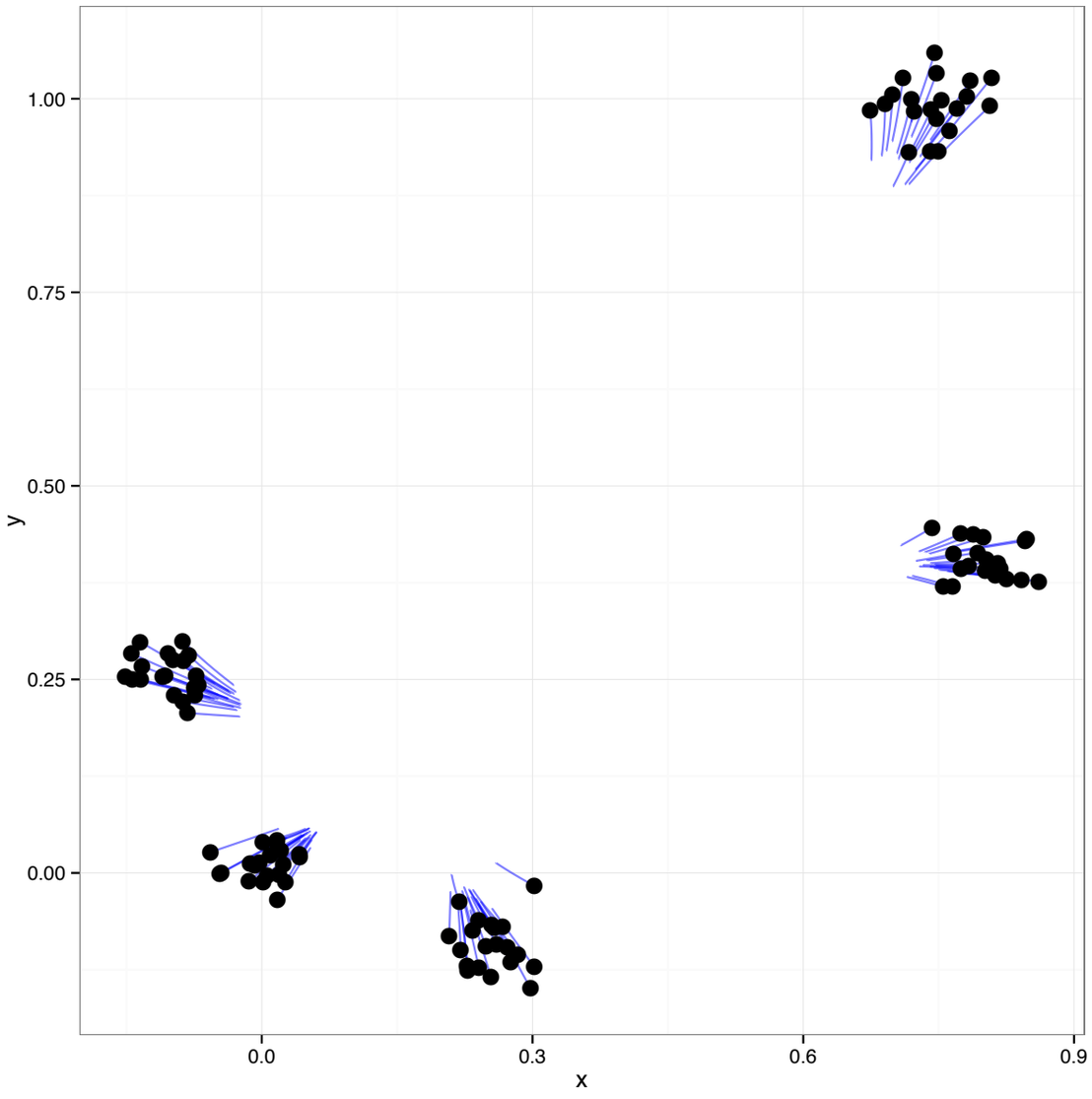
The Solution Path



γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma\sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

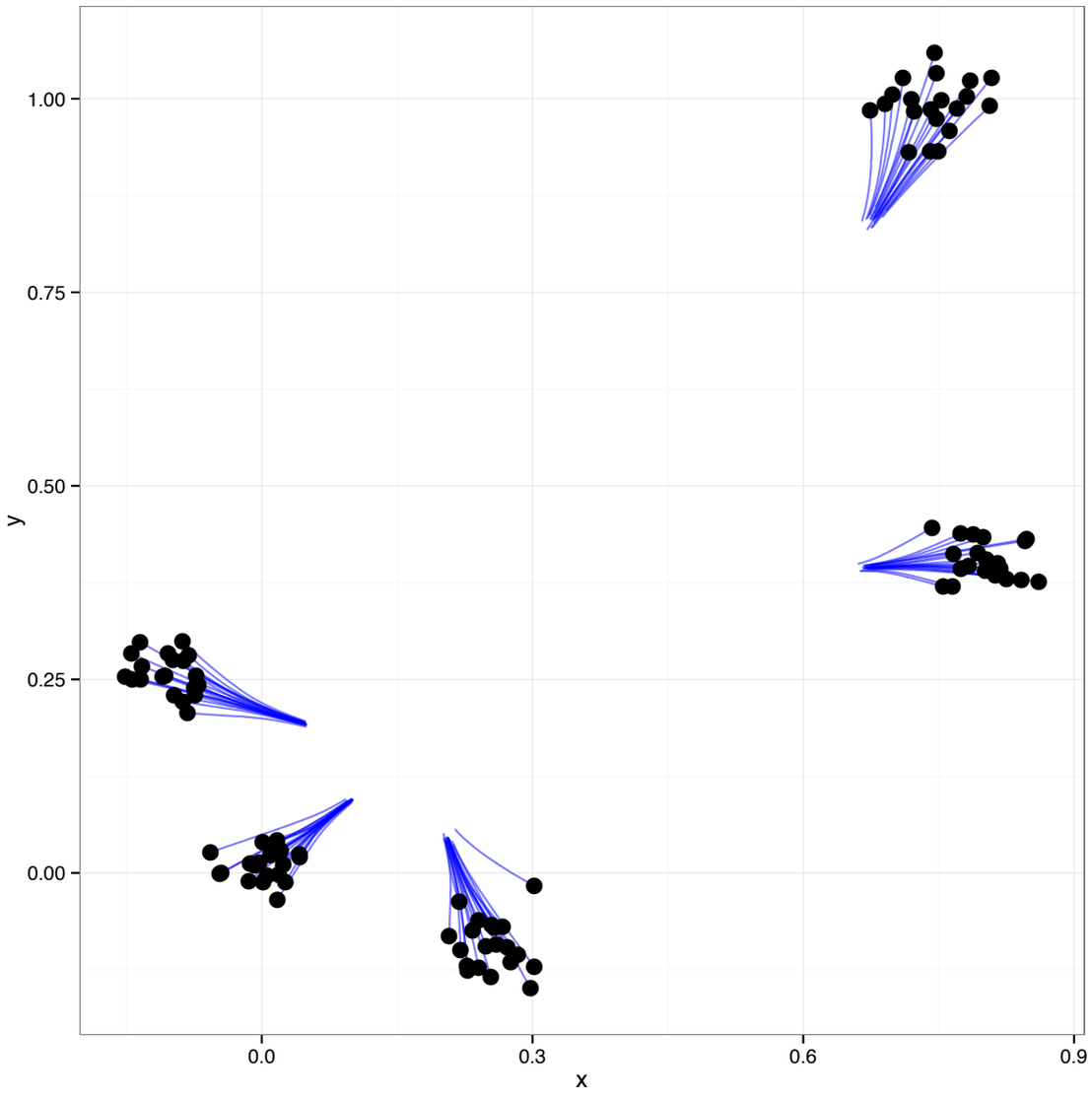
The Solution Path



γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

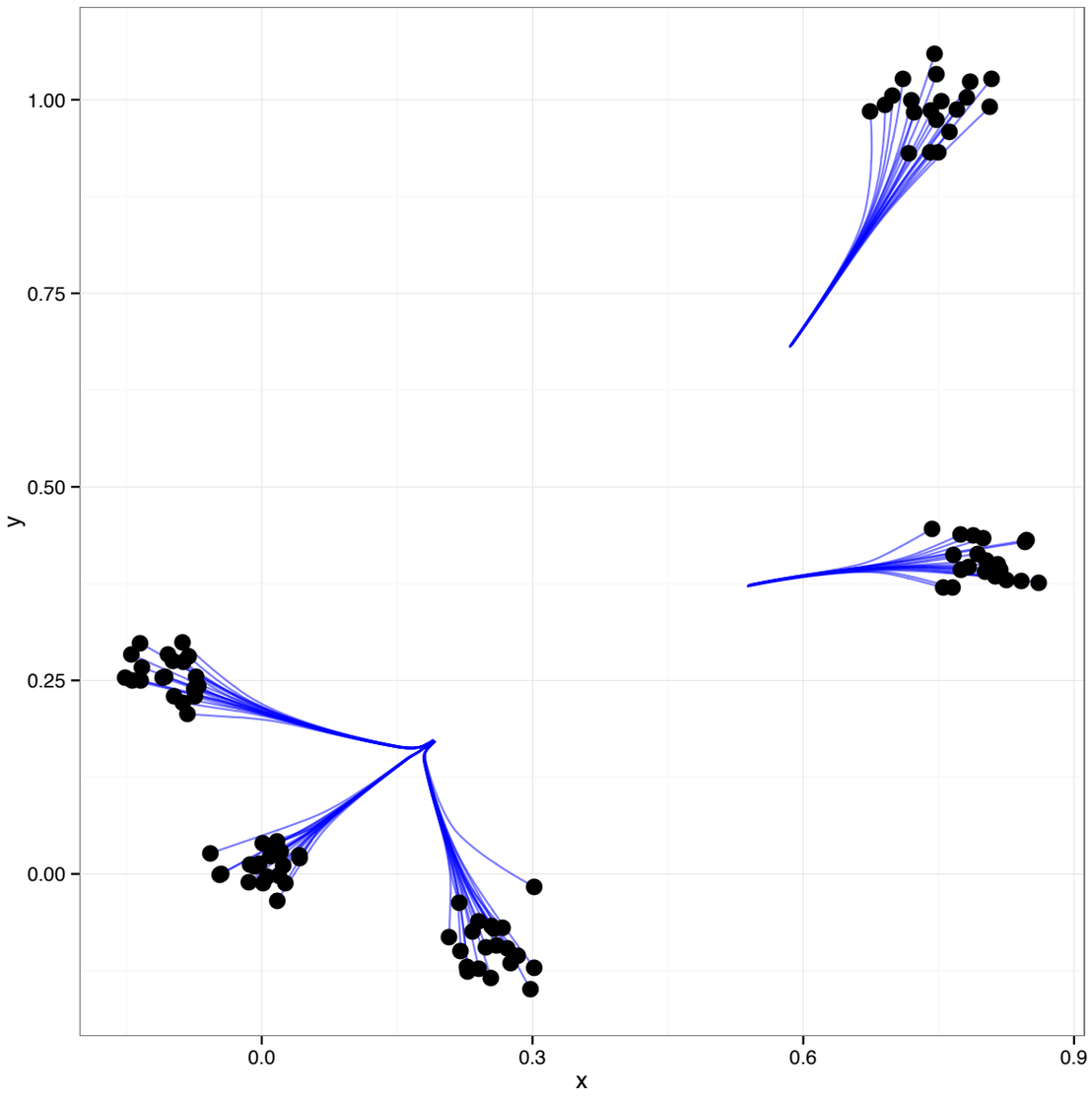
The Solution Path



γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma\sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

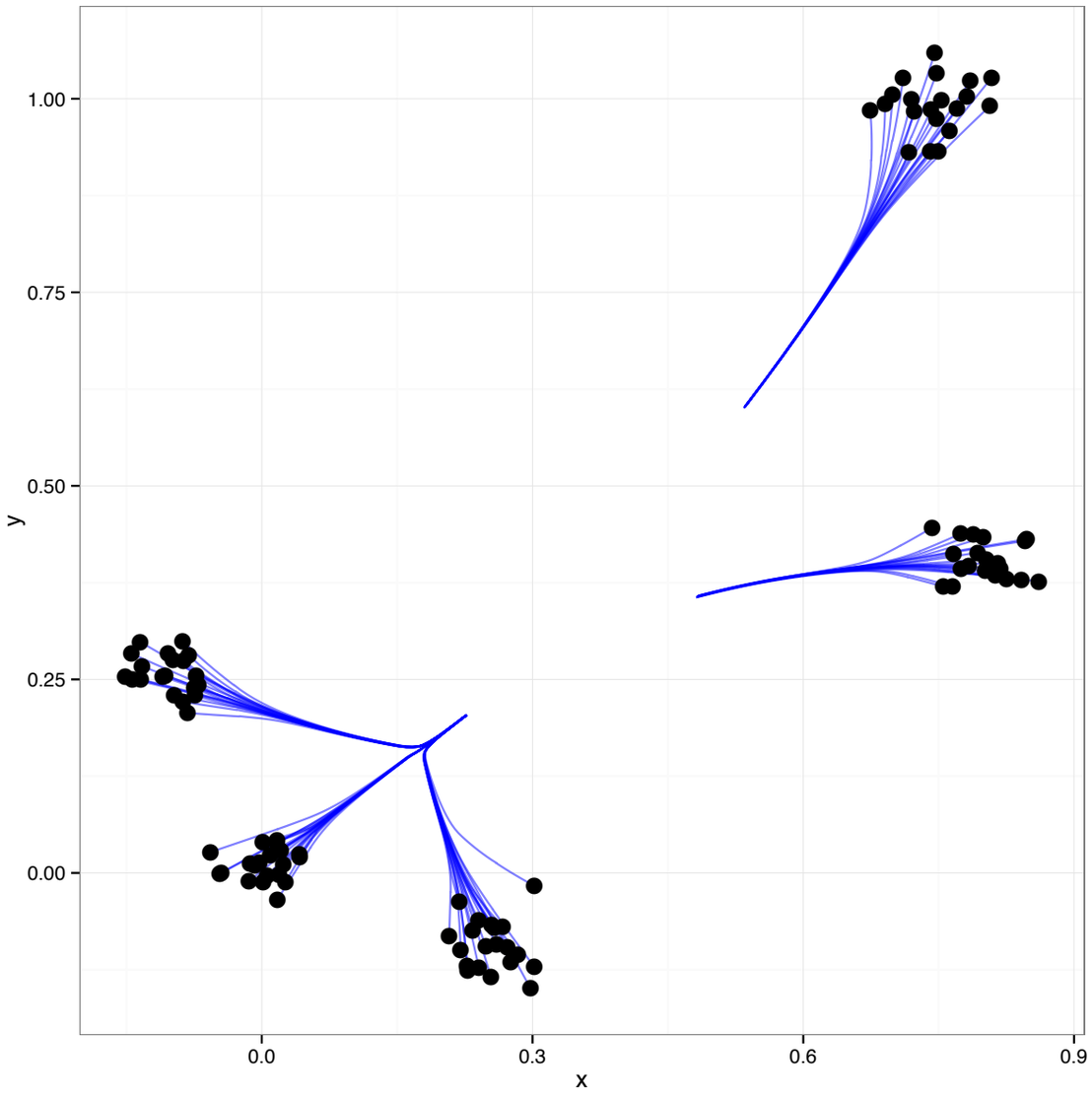
The Solution Path



γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

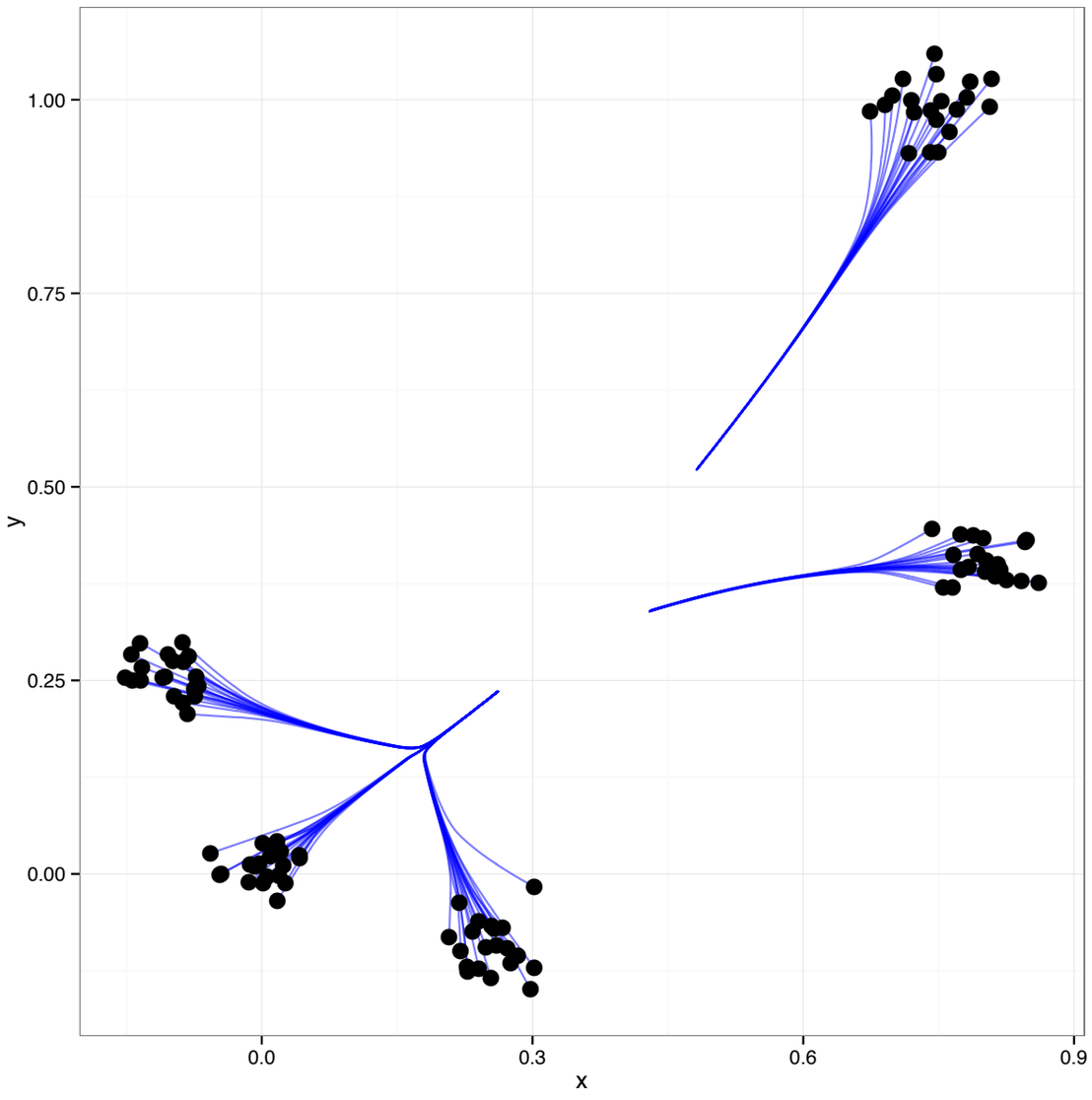
The Solution Path



γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

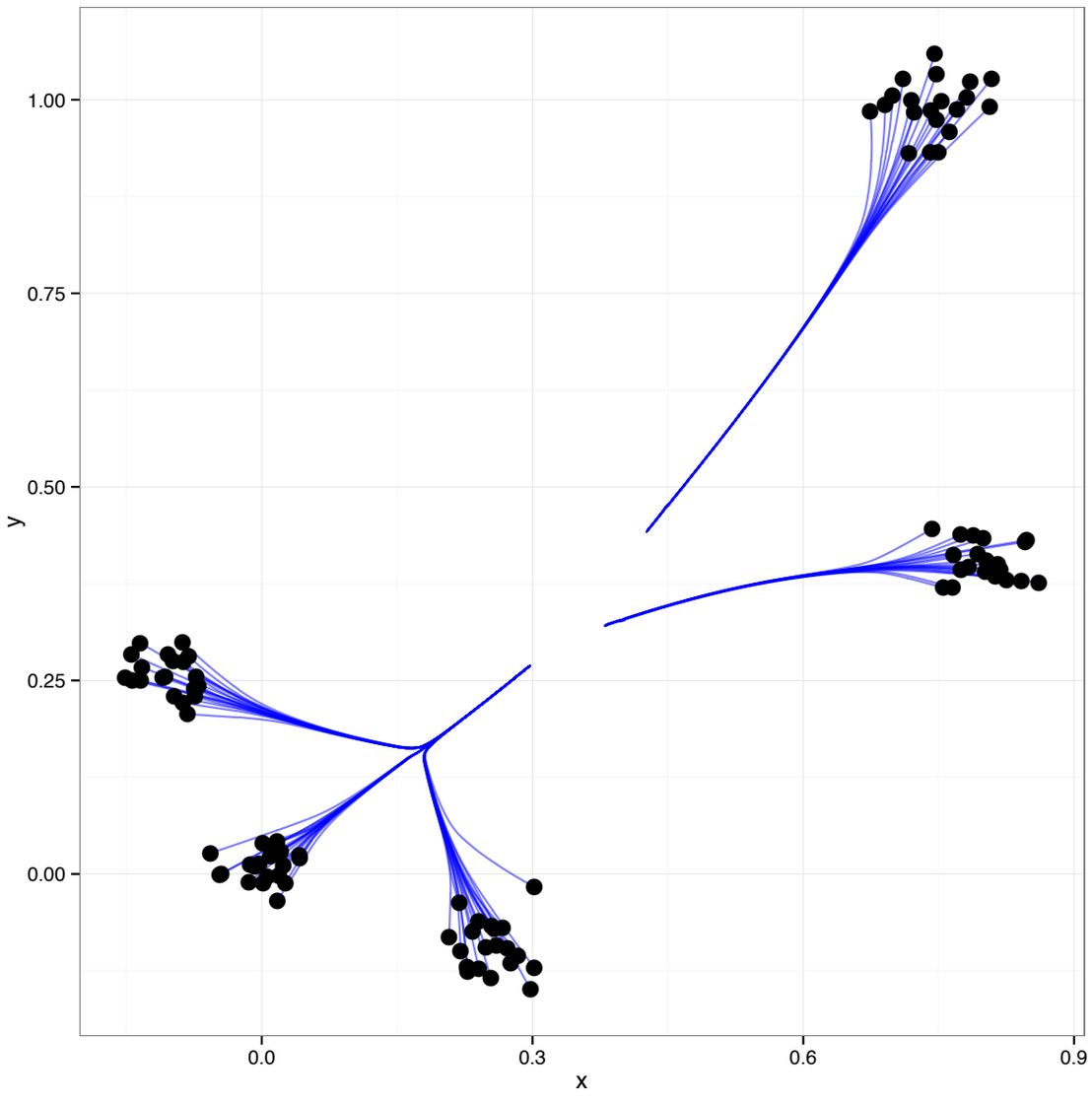
The Solution Path



γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

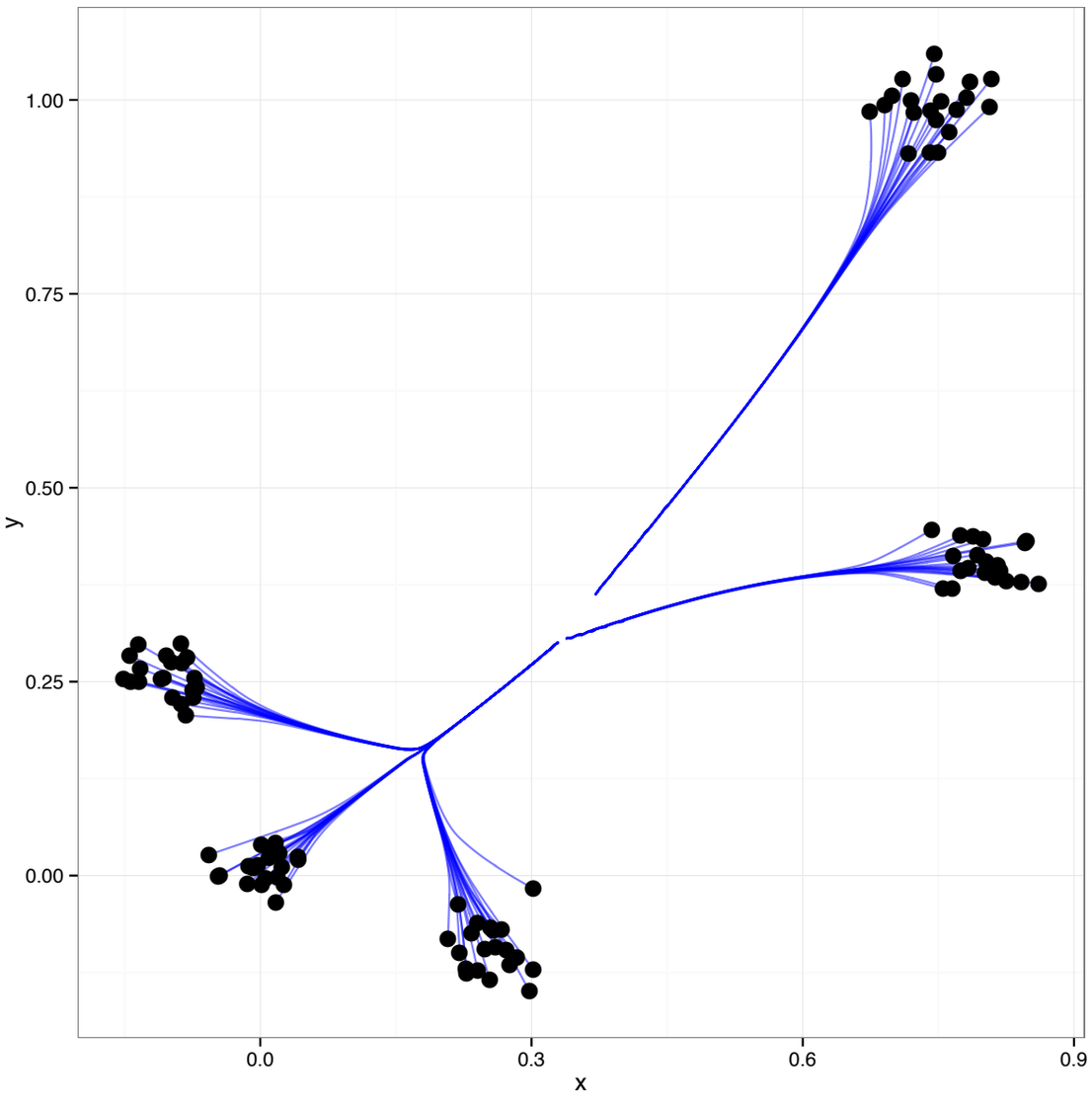
The Solution Path



γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

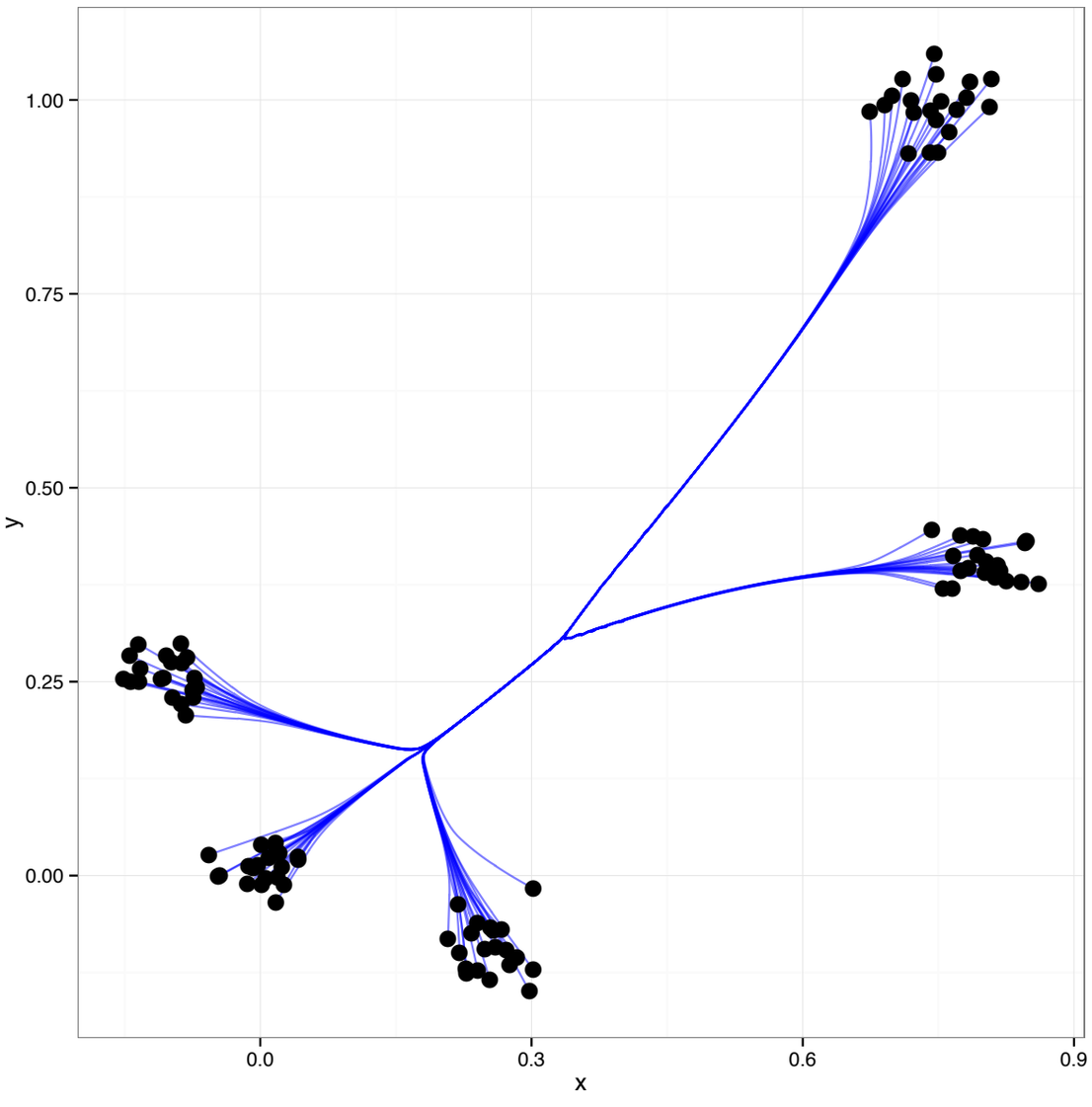
The Solution Path

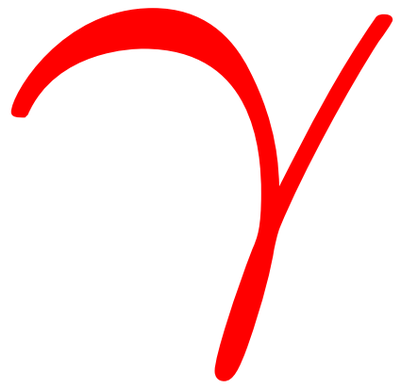


γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

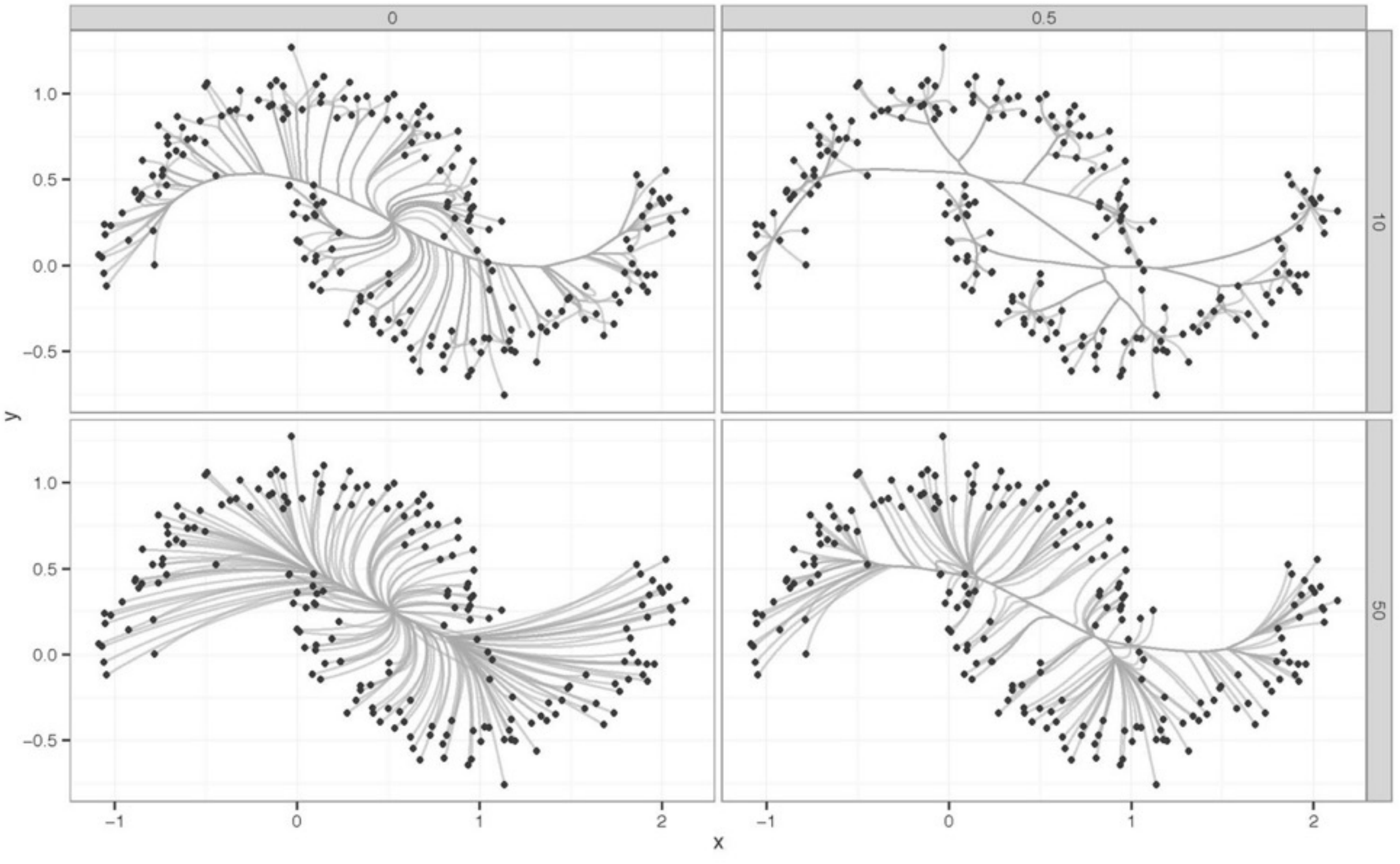
The Solution Path



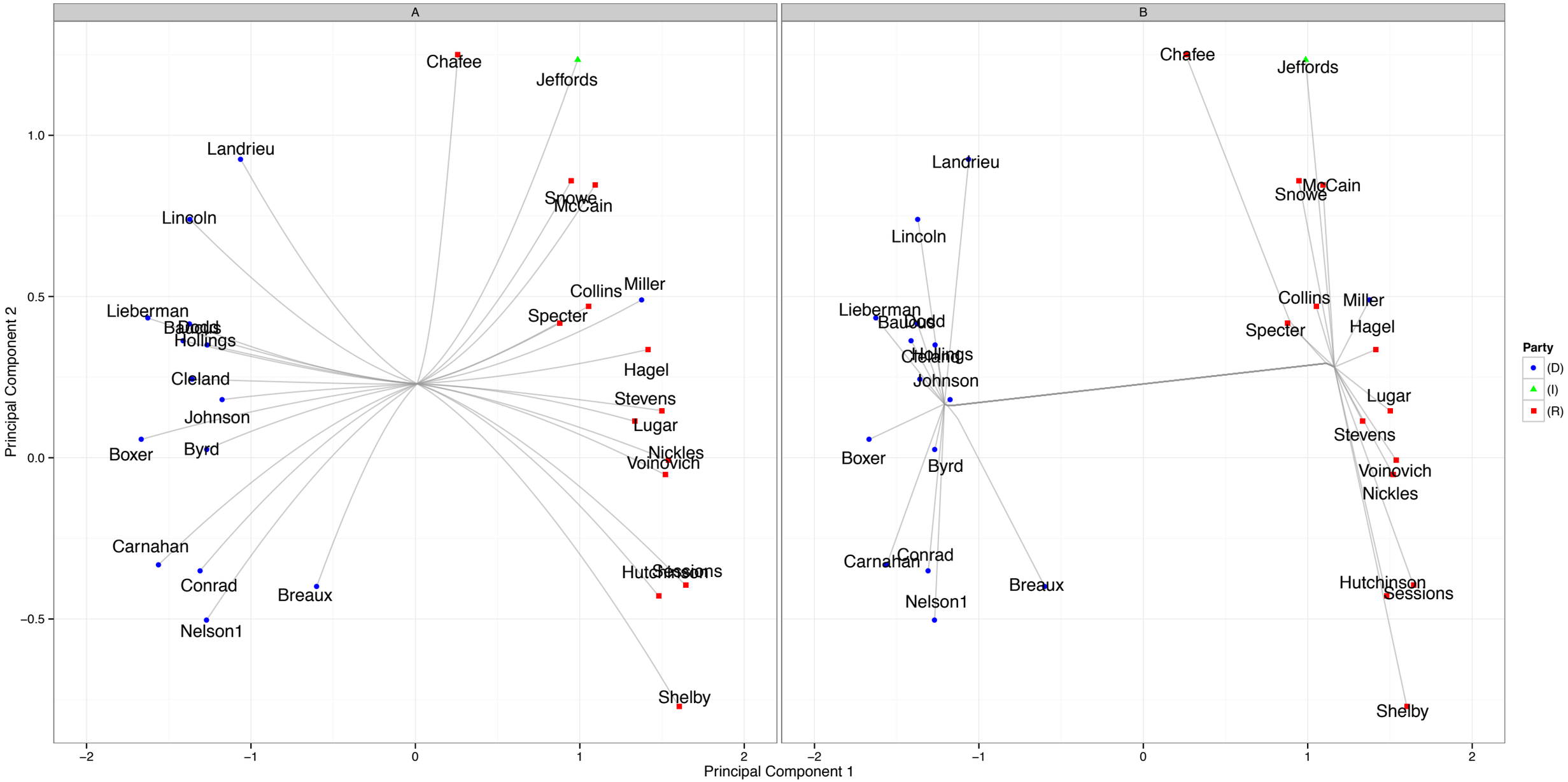


$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma\sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

Two Interlocking Half-Moons



Senate Voting



Apparently Non-Trivial Optimization Problem

Why is this hard to solve?

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

Apparently Non-Trivial Optimization Problem

Why is this hard to solve?

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

Nonsmooth? Not the issue

Apparently Non-Trivial Optimization Problem

Why is this hard to solve?

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

Affine transformation of u

Convex Clustering: Variable Split Version

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{p}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma\sum_{l} w_l\|\mathbf{v}_l\|$$

$$\text{subject to } \mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \mathbf{v}_l = \mathbf{0}$$

$$l = (l_1, l_2) \text{ with } l_1 < l_2.$$

ADMM Updates

$$
\begin{aligned}
\mathbf{u}_i &= \frac{1}{1+p\nu}\mathbf{y}_i + \frac{p\nu}{1+p\nu}\bar{\mathbf{x}} \\
\mathbf{y}_i &= \mathbf{x}_i + \sum_{l_1=i}[\boldsymbol{\lambda}_l + \nu\mathbf{v}_l] - \sum_{l_2=i}[\boldsymbol{\lambda}_l + \nu\mathbf{v}_l].
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{v}_l &= \arg\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{v} - (\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l)\|_2^2 + \frac{\gamma w_l}{\nu}\|\mathbf{v}\| \\
&= \mathrm{prox}_{\sigma_l\|\cdot\|/\nu}(\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l),
\end{aligned}
$$

where $\sigma_l = \gamma w_l$.

$$
\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l + \nu(\mathbf{v}_l - \mathbf{u}_{l_1} + \mathbf{u}_{l_2}).
$$

AMA Updates

$$
\begin{aligned}
\mathbf{u}_i &= \frac{1}{1+p0}\mathbf{y}_i + \frac{p0}{1+p0}\bar{\mathbf{x}} \\
\mathbf{y}_i &= \mathbf{x}_i + \sum_{l_1=i}[\boldsymbol{\lambda}_l + 0\mathbf{v}_l] - \sum_{l_2=i}[\boldsymbol{\lambda}_l + 0\mathbf{v}_l].
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{v}_l &= \arg\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{v} - (\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l)\|_2^2 + \frac{\gamma w_l}{\nu}\|\mathbf{v}\| \\
&= \mathrm{prox}_{\sigma_l\|\cdot\|/\nu}(\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l),
\end{aligned}
$$

where $\sigma_l = \gamma w_l$.

$$
\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l + \nu(\mathbf{v}_l - \mathbf{u}_{l_1} + \mathbf{u}_{l_2}).
$$

AMA Updates

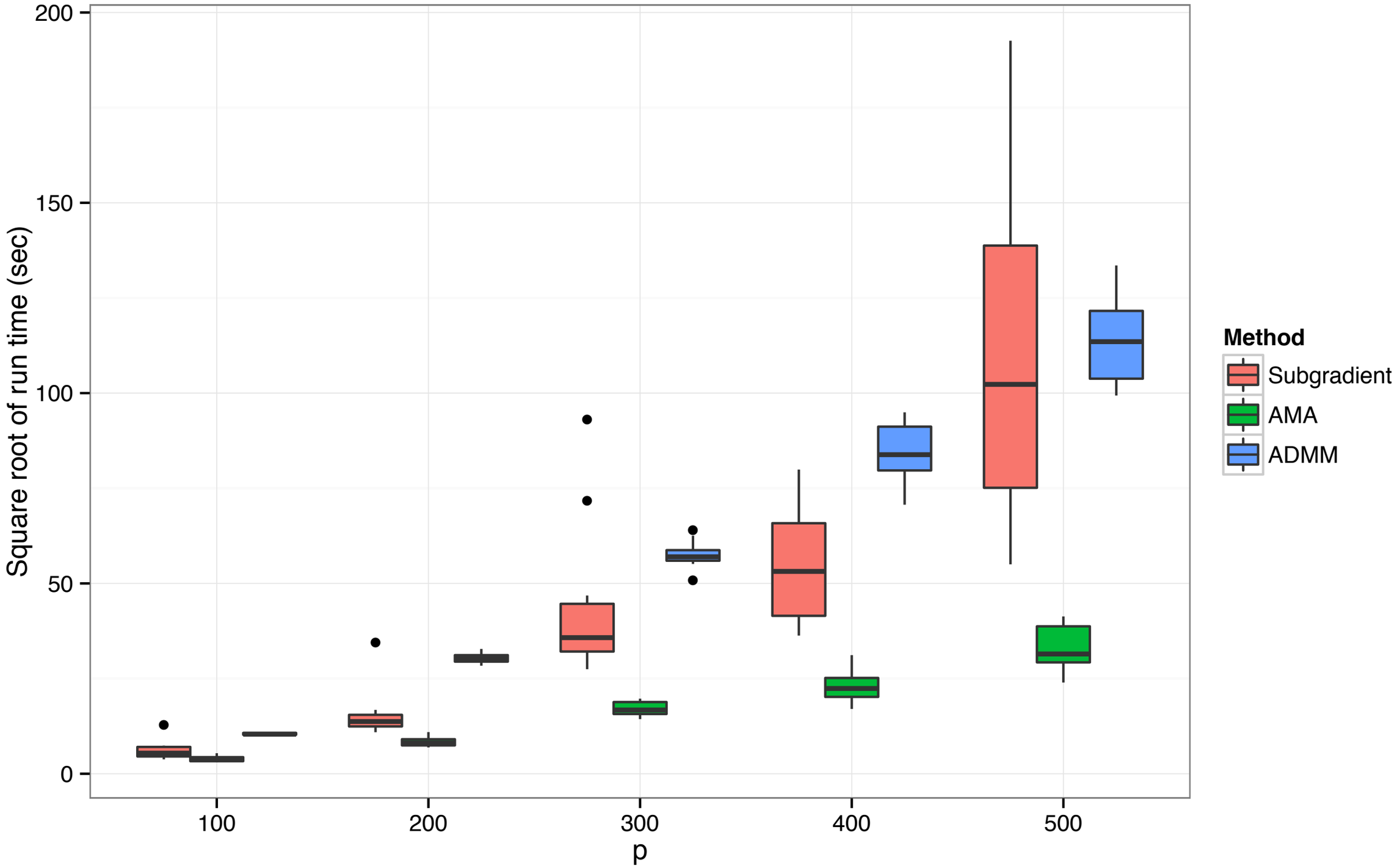
$$\mathbf{u}_i \quad = \quad \mathbf{x}_i + \sum_{l_1=i} \boldsymbol{\lambda}_l - \sum_{l_2=i} \boldsymbol{\lambda}_l$$

$$\begin{aligned}
\mathbf{v}_l &= \arg\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{v} - (\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l)\|_2^2 + \frac{\gamma w_l}{\nu}\|\mathbf{v}\| \\
&= \mathrm{prox}_{\sigma_l\|\cdot\|/\nu}(\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l),
\end{aligned}$$

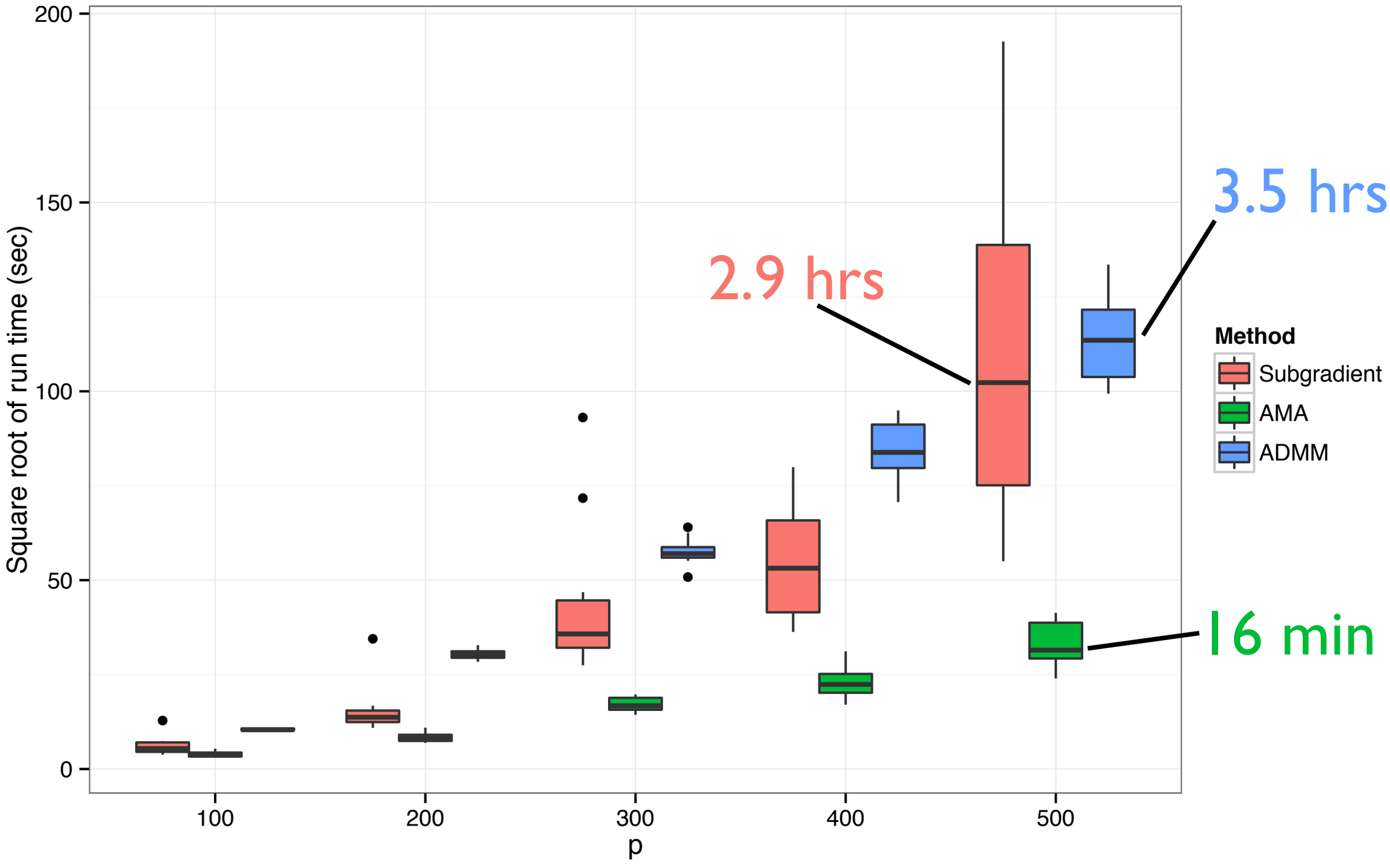
where $\sigma_l = \gamma w_l$.

$$\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l + \nu(\mathbf{v}_l - \mathbf{u}_{l_1} + \mathbf{u}_{l_2}).$$

What's the Difference?



What's the Difference?



ADMM versus AMA

$$\text{minimize } f(x) + g(y)$$
$$\text{subject to } Ax + By = c$$

ADMM:

- Pros: weaker conditions (f and g need to be closed and convex, problem needs to have a solution).
- Cons: will be slower (still scalable)

AMA:

- Pros: fast and scalable
- Cons: stronger conditions (f needs to be strongly convex)
- Tseng, P. (1991), "Applications of a Splitting Algorithm to Decomposition in Convex Programming and Variational Inequalities," SIAM Journal on Control and Optimization, 29, 119-138.

Summary

Duality is useful for

1. Checking correctness of a solution & debugging your code
2. Designing an alternative algorithm for solving an equivalent dual problem.