Smoothing Techniques in Constrained Stochastic Gradient Descent

Robert Bassett

January 7, 2015

The Problem

We want to solve the problem

$$\min \sum_{i=1}^{k} f_i(x)$$

$$\text{subject to } Ax \leq b$$

where

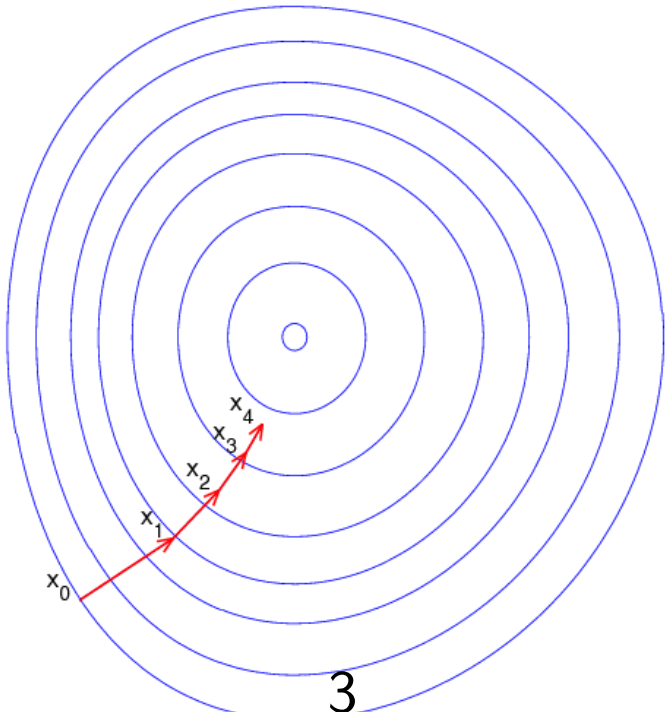
- k is very large
- $x \in \mathbf{R}^n$,
- A is $m \times n$
- Each of the f_i are convex, i.e. $\alpha \in [0, 1]$ gives

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

The Context

In the unconstrained case, when k is not unreasonably gigantic we would use the *Gradient Descent Method*.

- Gradient descent proceeds by moving in the direction of $-\sum_{i=1}^{k} \nabla f_i(x_n)$ at each step.
- Nesterov (1983) showed error in step n is $O(\frac{1}{n^2})$.



The Context

In the unconstrained case when k is gigantic, computing the full gradient can be computationally expensive. Instead we proceed with the following, where γ is suitable choice of step length.

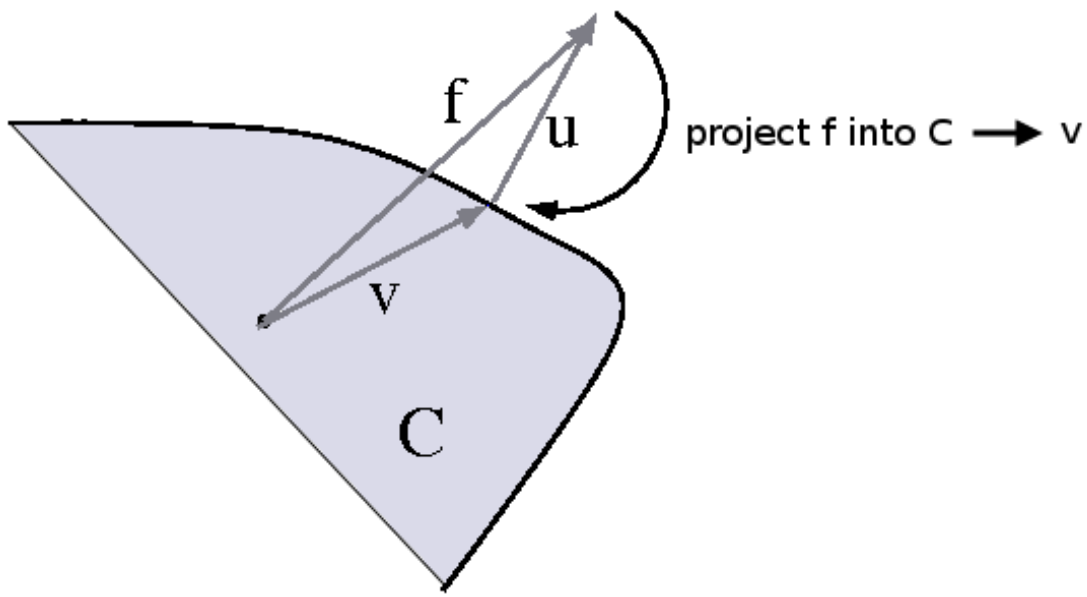
- $x_0 = 0$
- Choose $j \in \{1, ..., k\}$ at random (uniformly)
- $x_{n+1} \leftarrow x_n - \gamma_n \nabla f_j(x_n)$
- Repeat steps 2 and 3.

Fact of Life: This algorithm converges almost surely to a global minimum.

The Context

But we have constraints!

- A common way to deal with this is to run SGD and project back into the feasible region at each iteration.
- Projection can be expensive! This is a bottleneck on the algorithm.

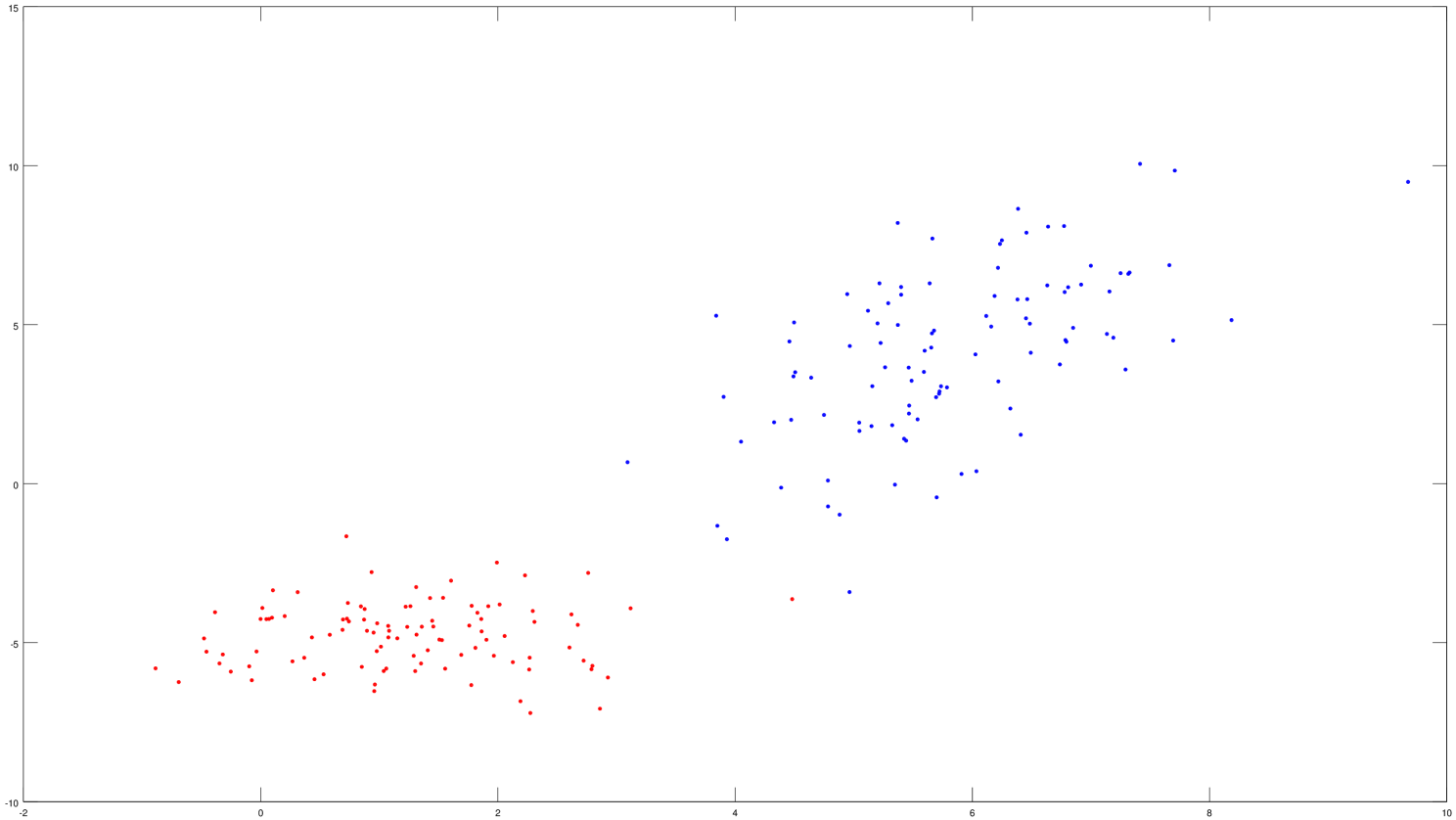


The Question

How can we perform constrained Stochastic Gradient Descent while avoiding computationally expensive projections?

A Running Example

In the linear separation problem, we have two clusters of points and seek to find the "best" line that separates them.



A Running Example

This is often done by solving the least squares problem

$$\min ||Y - [1\ X]\beta||^2$$

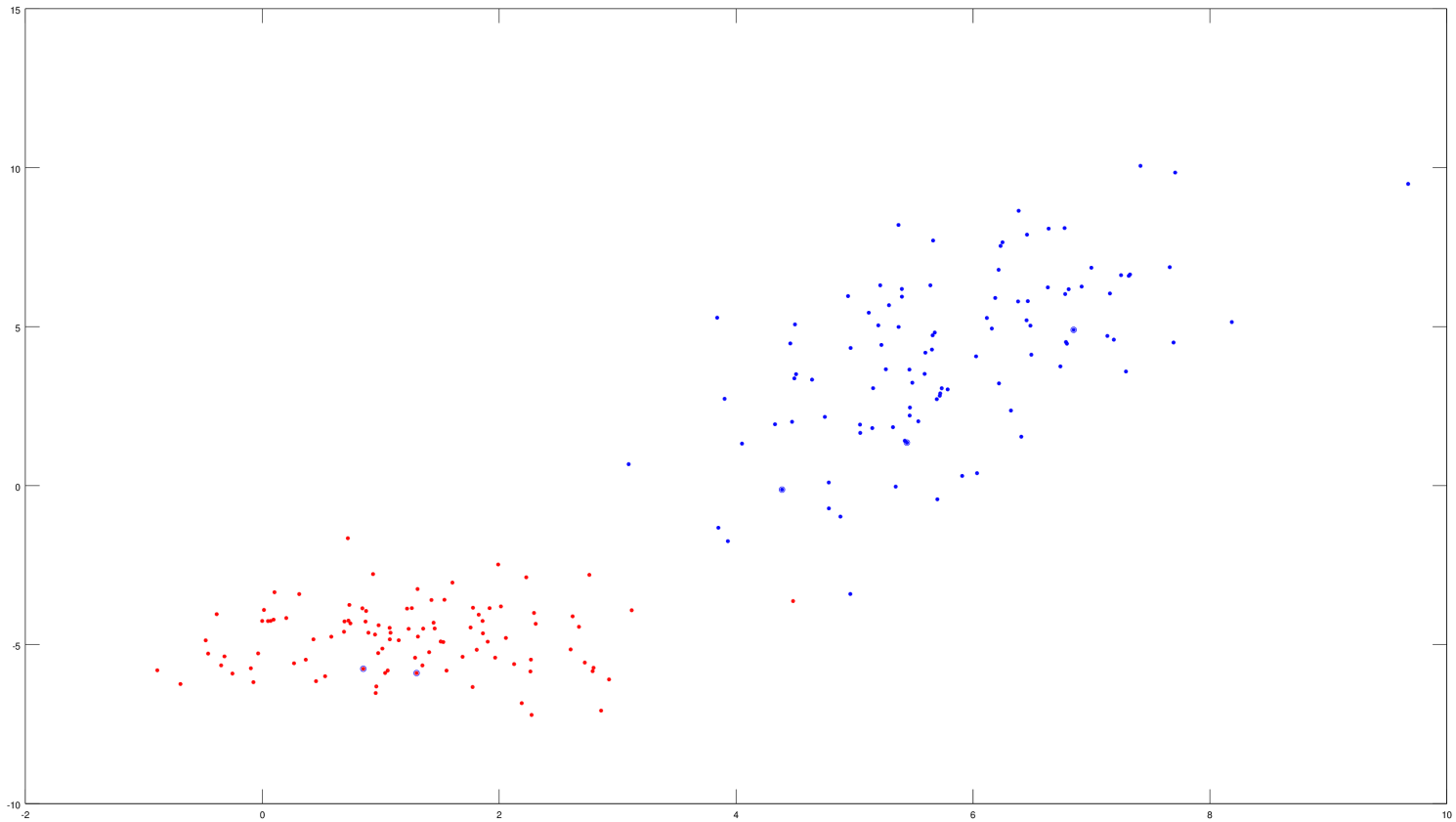
for beta.

- X is a matrix whose rows are made up of all points in the problem.
- Y is a column vector. $Y_i = 1$ if X_i is blue, and -1 if X_i is red.
- *beta* is a column vector that describes a hyperplane

$$\beta_0 + \beta_1 x_1 + ... + \beta_n x_n = 0$$

A Running Example

But we need constraints! In our example, certain points are chosen at random beforehand as "preclassified", i.e. we must have them on a certain side of the linear separator.



A Running Example

Our example is

$$\min ||Y - [1\ X]\beta||^2$$

$$\text{subject to } [1\ X_s]\beta \leq 0 \text{ and } [1\ X_{s'}]\beta \geq 0$$

which is of the form

$$\sum_{i=1}^{k} f_i(\beta)$$

$$\text{subject to } A\beta \leq 0$$

for

$$f_i(\beta) = (Y_i - [1\ X_i]\beta)^2$$

and

$$A = \begin{bmatrix} 1 & X_s \\ -1 & -X_{s'} \end{bmatrix}.$$

So this fits our framework.

Indicator Functions

An important tool at our disposal is the notion of an indicator function

Definition

The *indicator function* of a set $C \subset \mathbf{R}^n$ is $\delta(\cdot|C) : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ with

$$\delta(x|C) = \begin{cases} 0 & \text{if} \quad x \in C \\ +\infty & \text{if} \quad x \notin C \end{cases}$$

Indicator functions have the following nice properties

- $\delta(\cdot|C)$ is convex if and only if C is convex
- $\delta(\cdot|C)$ is lower-semicontinuous ($x_n \to x \Rightarrow f(x) \leq \liminf f(x_n)$) if and only if C is closed.

Indicator Functions

Let C be the set $\{\beta : A\beta \leq 0\}$ where A is the same as in our running example. Then

$$\min ||Y - [1\ X]\beta||^2$$

$$\text{subject to } A\beta \leq 0$$

and

$$\min ||Y - [1\ X]\beta||^2 + \delta(\beta|C)$$

are equivalent problems.

We have moved the constraint into the objective function, but have sacrificed smoothness and made use of the extended real line.

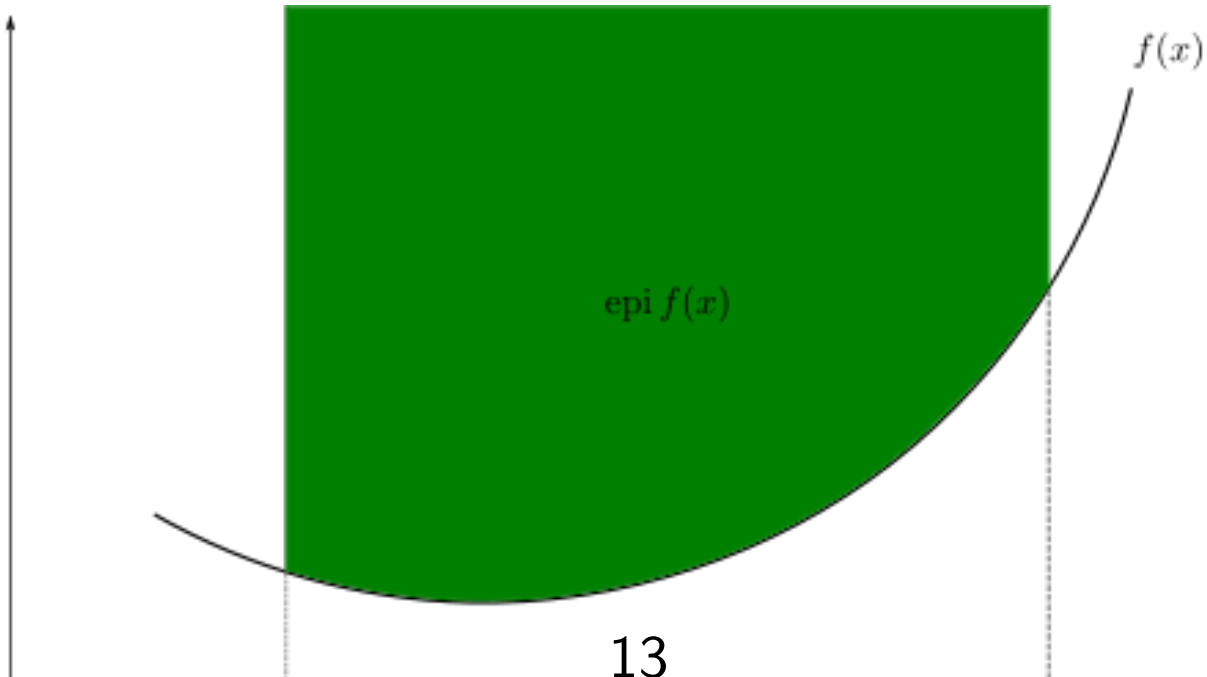
Epigraph

We will now move to regain the smoothness lost in the previous step with the hope of applying SGD.

Definition

For an extended real-valued function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ its epigraph is given by

$$\text{epi } f := \{(x, \alpha) \in R^n \times \mathbf{R} | f(x) \leq \alpha\}$$



Epi-convergence

Definition

We say that a sequence $\{f_k\}$ of functions $f_k : \mathbf{R}^n \to \overline{R}$ *epi-converges* to $f : \mathbf{R}^n \to \overline{R}$ if

$$\lim_{k\to\infty} \text{epi } f_k = \text{epi } f,$$

where a Panleve-Kuratowski notion of set-convergence is employed.

In this case we write $f_k \xrightarrow{e} f$.

A handy characterization is the following:

$$f_k \xrightarrow{e} f \iff \forall \overline{x} \in \mathbf{R}^n \begin{cases} \forall x_n \to \overline{x} : \ \liminf f_k(x_n) \le f(\overline{x}), \\ \exists x_n \to \overline{x} : \limsup f_k(x_n) \le f(\overline{x}) \end{cases}$$

Epi-convergence

Why is epi-convergence important? The following theorem justifies its use

Definition

A function f is *level-bounded* if for every α the set $\{x \in \mathbf{R}^n | f(x) \leq \alpha\}$ is bounded.

Theorem

Suppose the sequence $\{f^\nu\}$, $\nu \in \mathbb{N}$ *is eventually level-bounded, and* $f^\nu \xrightarrow{e} f$ *with* f^ν *and* f *lsc. Then*

$$\inf f^\nu \to \inf f$$

and

$$\limsup_\nu (\mathit{argmin}\ f^\nu) \subset \mathit{argmin}\ f\,.$$

Epi-smoothers

The last few definitions and theorems give us a theory for constructing approximate problems via epi-convergent sequences of functions. Next we look to smoothness

Definition

Let $f : \mathbf{R}^n \to R \cup \{+\infty\}$ be lsc. We say $s_f : \mathbf{R}^n \times \mathbf{R}_+ \to \mathbf{R}$ is an *epi-smoothing function* for f if

- $s_f(\cdot, \mu_k)$ epi-converges to f for all $\{\mu_k\} \downarrow 0$.
- $s_f(\cdot, \mu)$ is continuously differentiable for all $\mu > 0$.

Summary

Pause for a deep breath!

So far we have

- Used indicator functions to encode constraints in an objective function.
- Defined an epigraph and the notion of epi-convergence of functions
- Seen that under mild conditions (level-boundedness and lsc) epi-convergence of functions gives convergence of minimizers and argmins.
- Defined an epi-smoother, which is essentially a smooth sequence of functions that epi-converges.

Summary

Next we will

- Look at a specific type of epi-smoother, the *Moreau Envelope*.
- Smooth indicator functions!
- APPLY it to our linear separation example.

Moreau Envelope

A prominent method of regularization is the Moreau Envelope.

Definition

Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be lsc. The *Moreau envelope* of f, also called the *Moreau-Yosida Regularization* is

$$e_\mu f(x) = \inf_w \{f(w) + \frac{1}{2\mu} ||w - x||^2\}.$$

What makes us think that this thing would be smooth?

Proof of Smoothness

Theorem

Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be lsc and convex, and bounded from below. Then the envelope function $e_\mu f$ is continuously differentiable

Proof of Smoothness

Theorem

Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be lsc and convex, and bounded from below. Then the envelope function $e_\mu f$ is continuously differentiable

Proof:

Fix x_0 in $\mathbf{R}^n$.

Since f is convex, the infimum in

$$e_\mu f(x_0) = \inf_w \{f(w) + \frac{1}{2\mu} ||w - x_0||^2\}$$

is unique. Call it $\overline{w}$.

To show differentiability at x_0, we need to show that there is a v such that

$$\lim_{u \to 0} \frac{|e_\mu f(x_0 + u) - e_\mu f(x_0) - \langle v, u \rangle|}{||u||}$$

Proof of Smoothness

Set $v = \frac{x_0 - \overline{w}}{\mu}$.

Consider

$$e_\mu f(x_0 + u) - e_\mu(x_0) - \langle v, u \rangle$$

$$= \inf_w f(w) + \frac{1}{2\mu} ||w - (x_0 + u)||^2 - f(\overline{w}) - \frac{1}{2\mu} ||\overline{w} - x_0||^2 - \langle v, u \rangle$$

$$\leq f(\overline{w}) + \frac{1}{2\mu} ||\overline{w} - (x_0 + u)||^2 - f(\overline{w}) - \frac{1}{2\mu} ||\overline{w} - x_0||^2 - \langle v, u \rangle$$

Proof of Smoothness

Expanding

$$= 2\frac{1}{2\mu}\langle \overline{w} - x_0, u\rangle + \frac{1}{2\mu}\left|\left|u\right|\right|^2 - \langle v, u\rangle$$

which, using the definition of v

$$= \frac{1}{2\mu}\left|\left|u\right|\right|^2.$$

So that

$$\lim_{u\to 0} \frac{|e_\mu f(x_0+u) - e_\mu(x_0) - \langle v, u\rangle|}{||u||} \leq \frac{||u||}{2\mu} \to 0.$$

And we're done!

Epi-convergence of Moreau Envelope

We also want that the Moreau Envelope of f epiconverges to f as $\mu \to 0$.

Theorem

Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be lsc and convex. The $e_\mu f(x)$ is an epi-smoothing function for f.

We will outsource this proof to Variational Analysis, Prop 7.4.

More Indicator Functions

Next we attempt to smooth indicator functions.

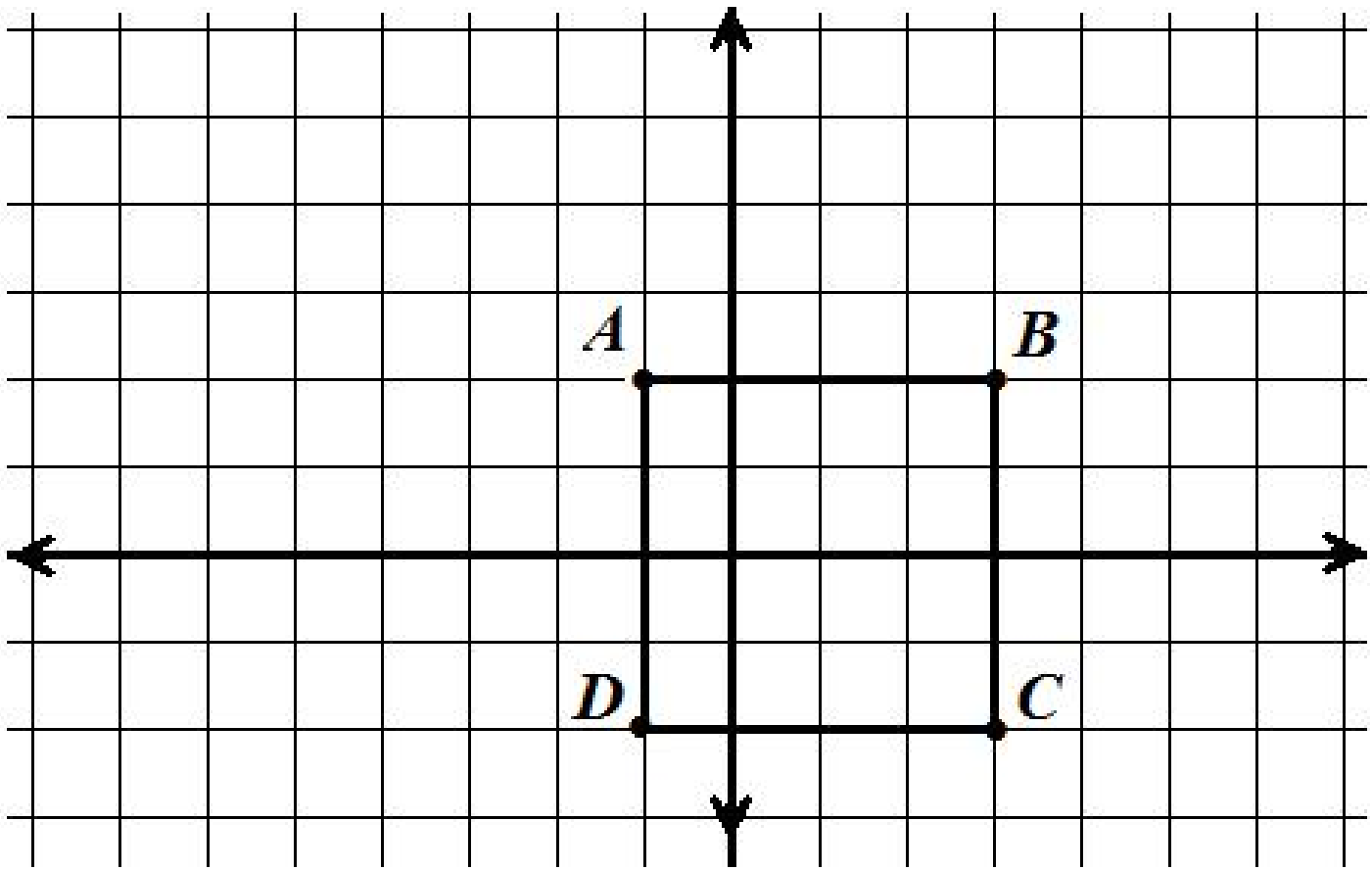
$$
\begin{aligned}
& e_\mu \delta(x|C) \\
&= \inf_w \delta(w|C) + \frac{1}{2\mu} ||w - x||^2 \\
&= \inf_{w \in C} \frac{1}{2\mu} ||w - x||^2 \\
&= \frac{1}{2\mu} ||Pr_C(x) - x||^2
\end{aligned}
$$

More projections? This is exactly what we were trying to avoid.

But we will attempt to exploit the fact that certain projections are easier to compute than others.

Exploiting Projections

If we are projecting onto a coordinate box, we can just take our point and project it into the box coordinate-wise. This is easy!



Exploiting Projections

In our example, we want to compute

$$e_\mu \delta(x|Ax \leq 0)$$

If we could ignore that A, we would have box constraints.

Exploiting Projections

In our example, we want to compute

$$e_\mu \delta(x|Ax \leq 0)$$

If we could ignore that A, we would have box constraints.

Theorem

If $A \in \mathbf{R}^{m \times n}$ has rank m and $b \in \mathbf{R}^m$, then $e_\mu \delta(Ax|x \leq 0)$ is an epi-smoothing function for $\delta(x|Ax \leq 0)$.

Problems?

But our A is usually tall and skinny, i.e. not rank m.

Problems?

But our A is usually tall and skinny, i.e. not rank m.

"This is the sort of restriction that I suspect does not often occur in practice." -Michael Friedlander

So let's try it anyways!

Implementation

Goal: Minimize the unconstrained problem

$$\min \quad \underbrace{||Y - [1\ X]\beta||^2}_{\text{Stochastic Gradient Descent}} \quad + \quad \underbrace{\delta(\beta|A\beta \leq 0)}_{\text{Smooth using Moreau Envelope}}$$

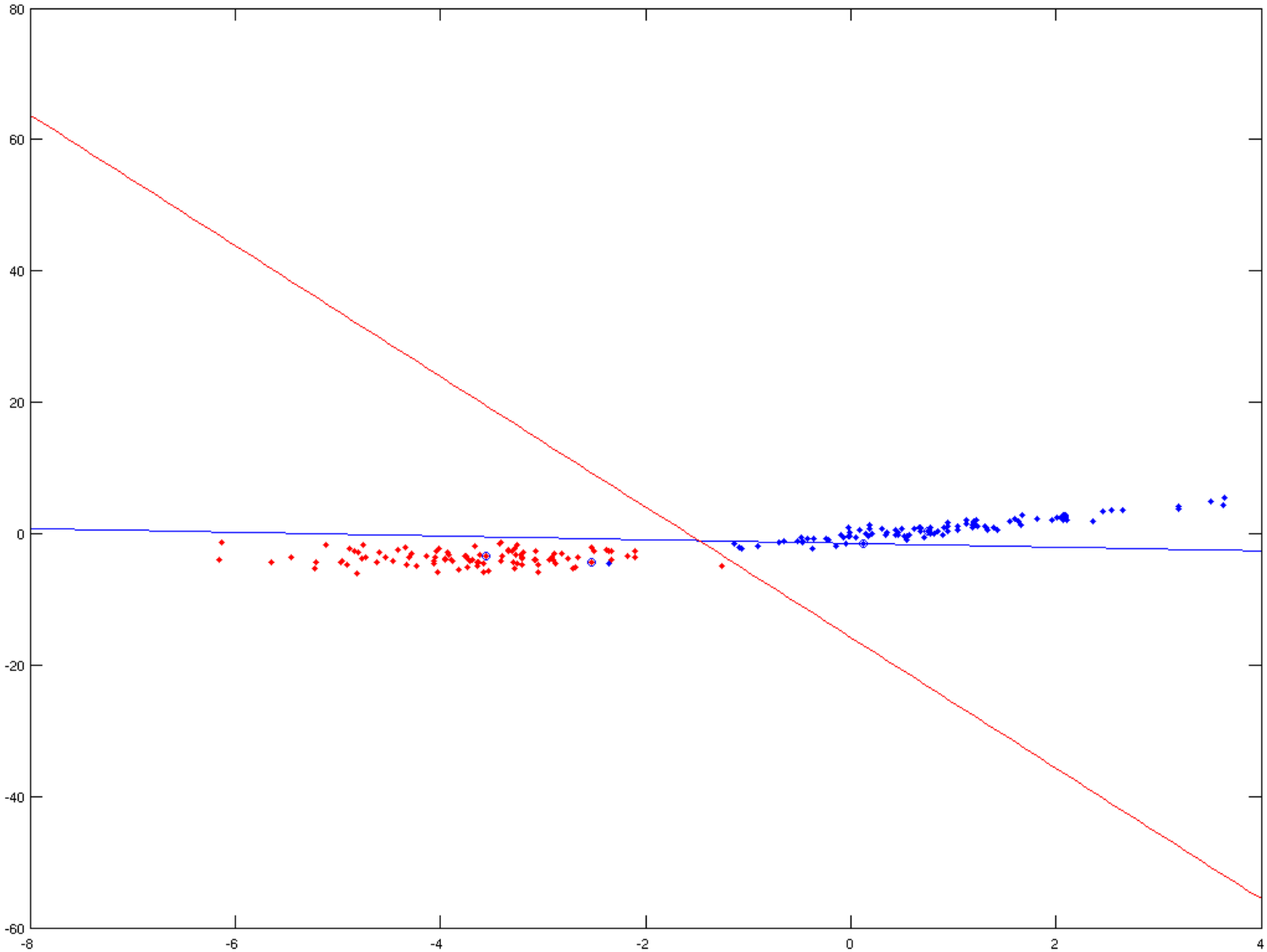
Details of implementation, where k is refers to iteration.

- $\gamma = (1.01)^{-k}$ (Stochastic step length)
- $\mu = (1.01)^{-k}$ (Smoothing Parameter)
- We applied SGD to the least squares term while applying GD to the smoothing term.

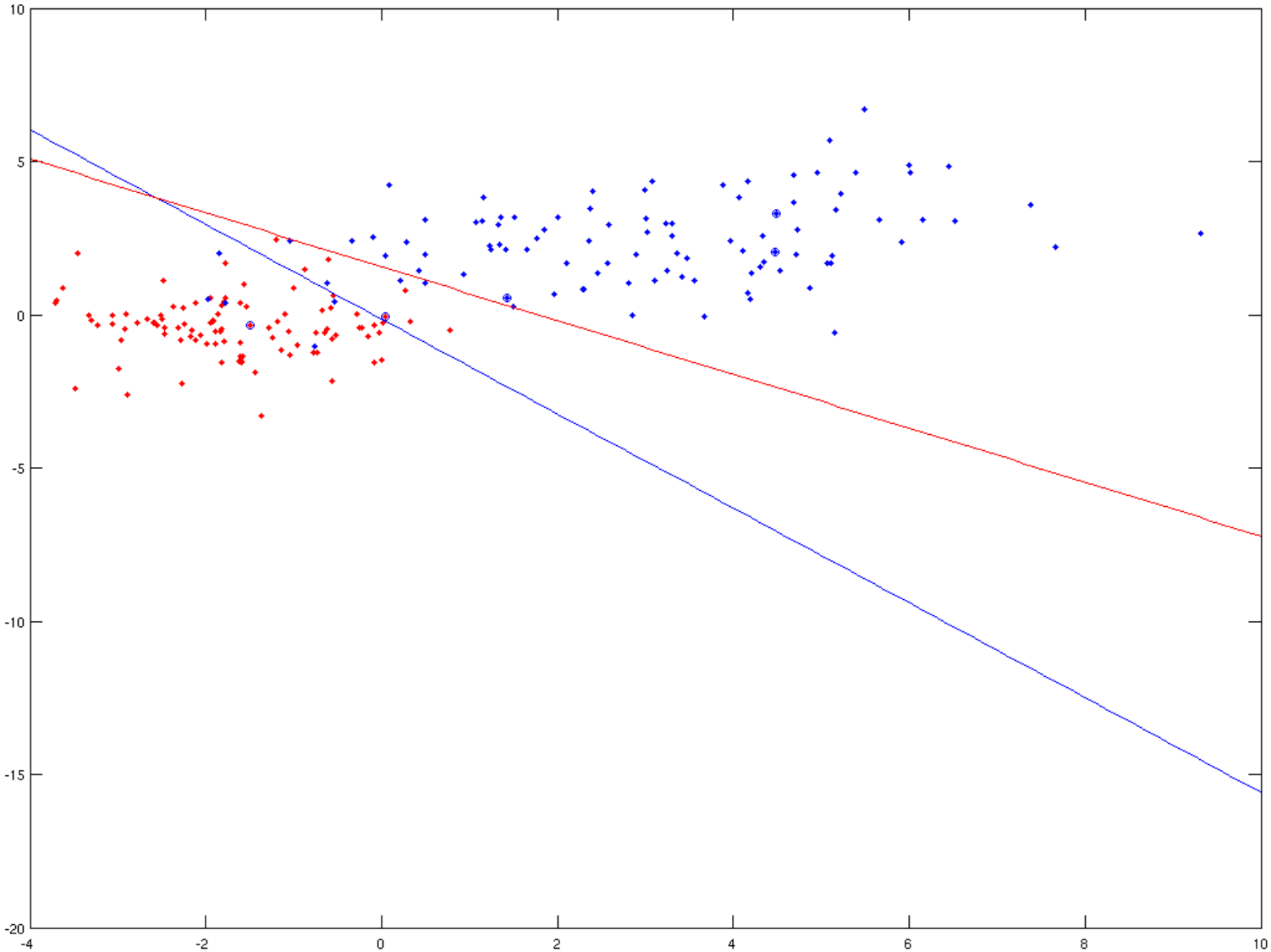
Results

- Green is unconstrained optimum (Least Squares)
- Red is constrained optimium (CVX)
- Blue is our method

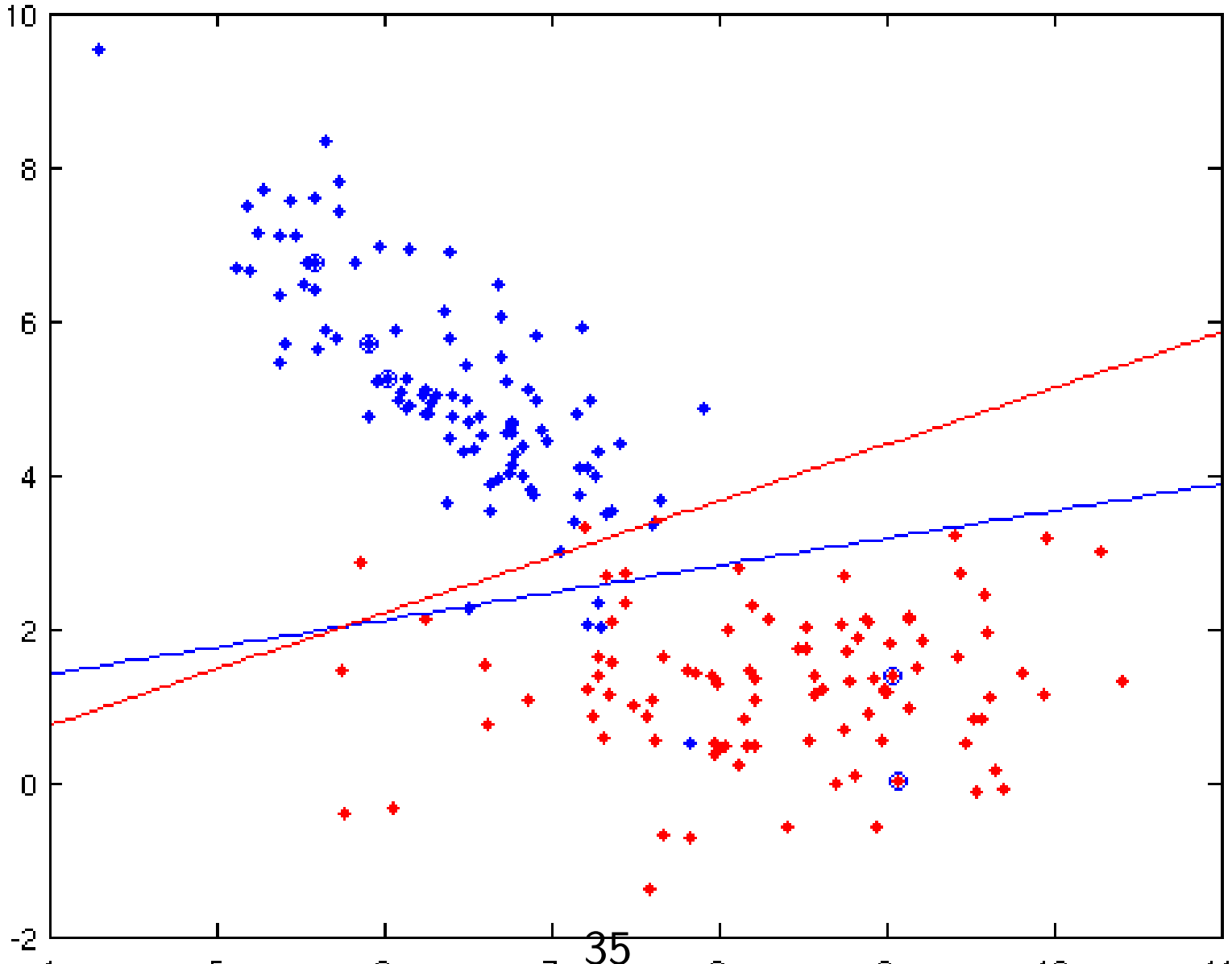
Results



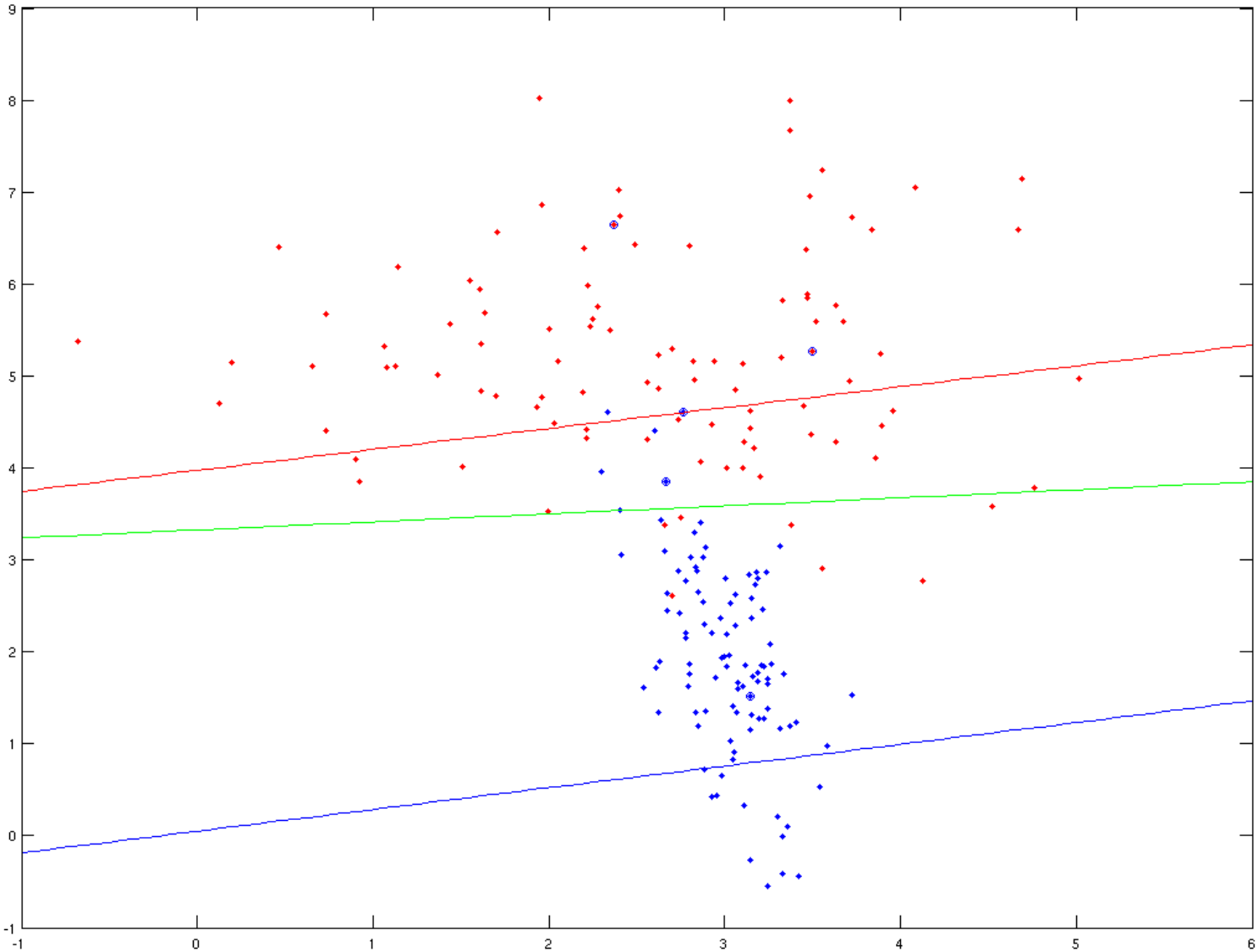
Results



Results



Results



References

- Beck, Teboulle. *Smoothing and First Order Methods: A Unified Framework*
- Rockafeller, Wets. *Variational Analysis*
- Burke, Hoheisel. *Epi-Convergent Smoothing with Applications to Convex Composite Functions*