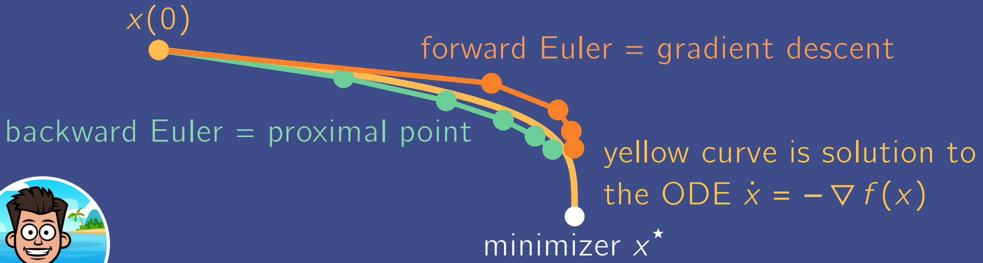
Continuous and Discrete Perspectives

How Euler's Methods relate to Gradient Descent and Proximal Point



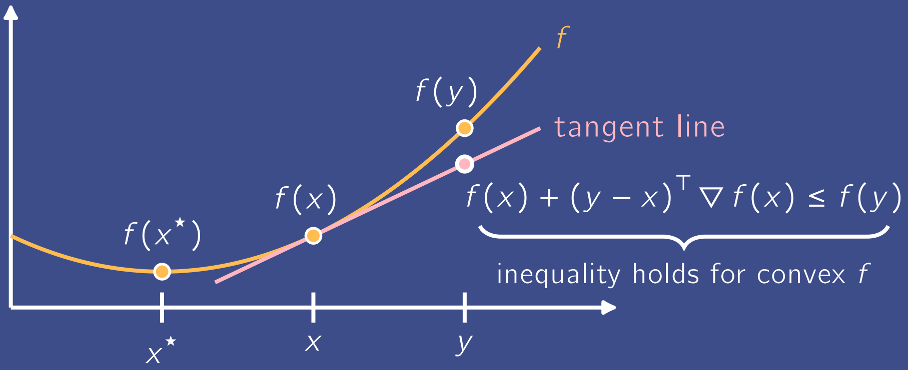
Howard Heaton

Typal Academy

Setting

For convex and differentiable $f(x)$ with minimizer $x^\star$, we consider the problem

$$\min_x f(x)$$



Outline

- Analyze paths of solutions to ordinary differential equation†

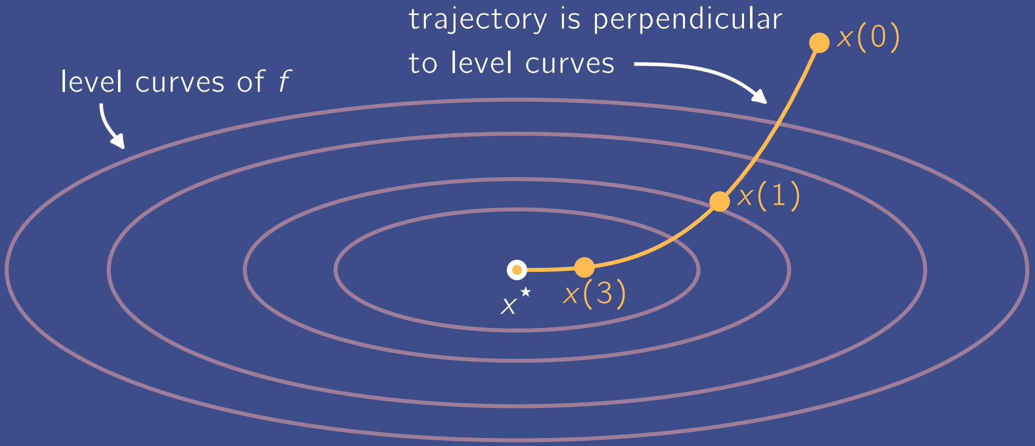
$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\nabla f(x)$$

- Relate forward Euler to gradient descent

- Relate backward Euler to proximal point

† The dependence of x on is t implicit, *i.e.* $x = x(t)$

Example Trajectory



As $t \to \infty$, trajectory $x(t)$ converges to minimizer $x^\star$ of $f(x) = x_1^2 + 3x_2^2$

Convergence Analysis

Consider the energy $\mathcal{E}(t)$ defined as the sum of two nonngative terms:

$$\mathcal{E}(t) = 2t\left[f(x) - f(x^\star)\right] + \|x - x^\star\|^2$$

This energy is monotonically decreasing (see next slide), which implies

$$f(x) - f(x^\star) \leq \frac{\mathcal{E}(t)}{2t} \leq \frac{\mathcal{E}(0)}{2t} = \frac{\|x - x^\star\|^2}{2t}$$

and so $f(x) \to f(x^\star) \quad \text{as} \quad t \to \infty$

Convergence Analysis

Differentiating the energy $\mathcal{E}(t)$ in time reveals[†]

$$\dot{\mathcal{E}} = 2\left[f(x) - f^\star\right] + 2t\nabla f(x)^\top \dot{x} + 2(x - x^\star)^\top \dot{x}$$

$$= \underbrace{2\left[f(x) + (x^\star - x)^\top \nabla f(x) - f^\star\right]}_{\leq 0 \text{ by convexity of } f} - \underbrace{2t\|\nabla f(x)\|^2}_{\leq 0}$$

$$\leq 0$$

Thus, $\dot{\mathcal{E}}(t) \leq 0$, and so $\mathcal{E}$ is monotonically decreasing

[†] Here we use dot notation for time derivatives, *i.e.* $\dot{x} = \mathrm{d}x/\mathrm{d}t$

Forward Euler

For time step $\lambda > 0$, set $x^k = x(k\lambda)$ so the forward Euler approximation is

$$\frac{x^{k+1} - x^k}{\lambda} = -\nabla f(x^k)$$

which may be rewritten as

$$x^{k+1} = x^k - \lambda \nabla f(x^k)$$

Gradient descent is forward Euler for our ODE

Backward Euler

The implicit approximation

$$\frac{x^{k+1} - x^k}{\lambda} = -\nabla f(x^{k+1})$$

may be written as

$$0 = \lambda \nabla f(x^{k+1}) + x^{k+1} - x^k$$

which holds when x^{k+1} solves

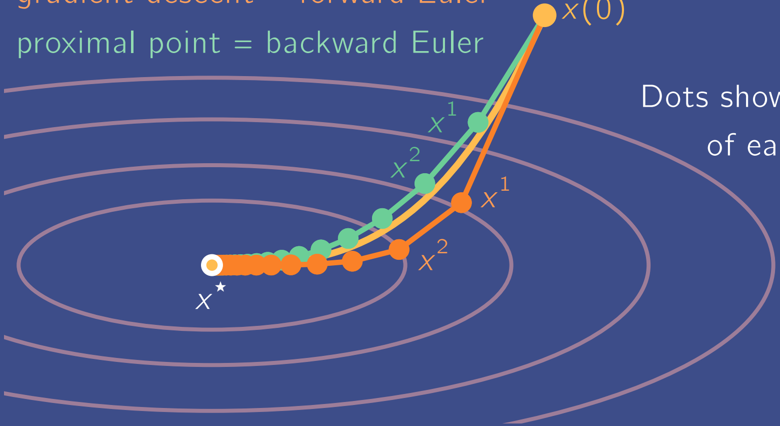
$$\min_x \lambda f(x) + \frac{1}{2}\|x - x^k\|^2$$

i.e. $x^{k+1} = \mathrm{prox}_{\lambda f}(x^k)$ and proximal point is backward Euler for our ODE

Example Trajectory

gradient descent = forward Euler

proximal point = backward Euler



Dots show discrete steps of each method

With appropriate λ, both forward Euler and backward Euler converge to $x^\star$

Takeaways

- Optimization algorithms typically have continuous analogues
- Continuous formulation is often simple to analyze
- Gradient descent and proximal point correspond to Euler discretizations

Appendix – Extensions

Proximal gradient for $f = g + h$ has both implicit and explicit terms:

$$\frac{x^{k+1} - x^k}{\lambda} \in -\partial g(x^{k+1}) - \nabla h(x^k)$$

$$\iff 0 \in \partial g(x^{k+1}) + \nabla h(x^k) + \frac{x^{k+1} - x^k}{\lambda}$$

$$\iff x^{k+1} = \operatorname*{argmin}_x g(x) + h(x^k) + \left\langle \nabla h(x^k), x - x^k \right\rangle + \frac{1}{2\lambda} \|x - x^k\|^2$$

$$\iff x^{k+1} = \operatorname{prox}_{\lambda g}(x^k - \nabla h(x^k))$$

Appendix – Extensions

- Runge-Kutta methods can be used to solve $\dot{x} = -\nabla f(x)$, but in optimization we are typically more interested in convergence to the limit $x^\star$ than matching this ODE trajectory

- Second-order ODEs can be discretized to give accelerated algorithms (*e.g.* Nesterov acceleration)