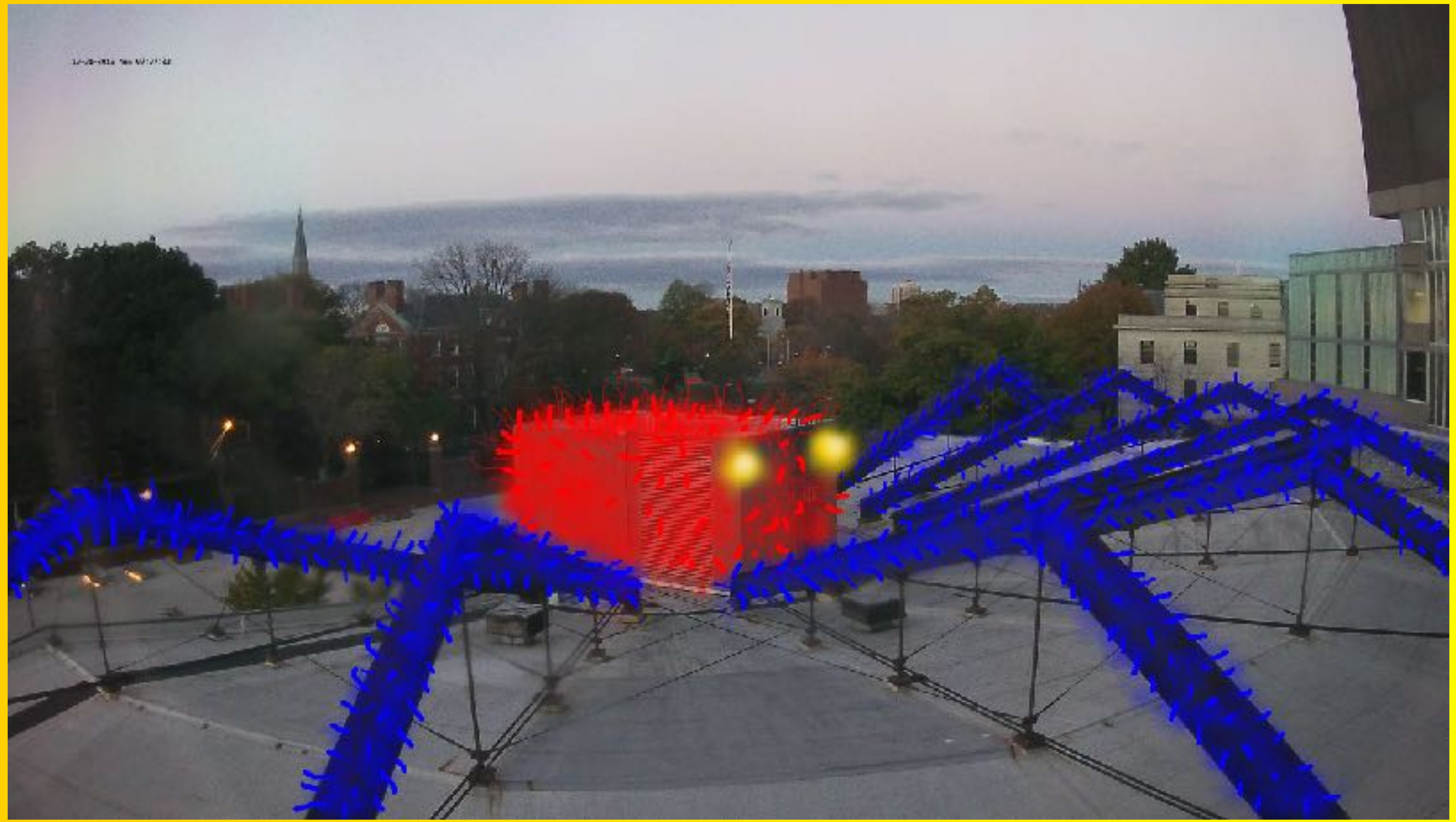


Lecture 24

Gradient Rule

Quadratic Approx

1) Gradient Theorem

2) Steepest Ascent

3) Quadratic approximation

4) Examples

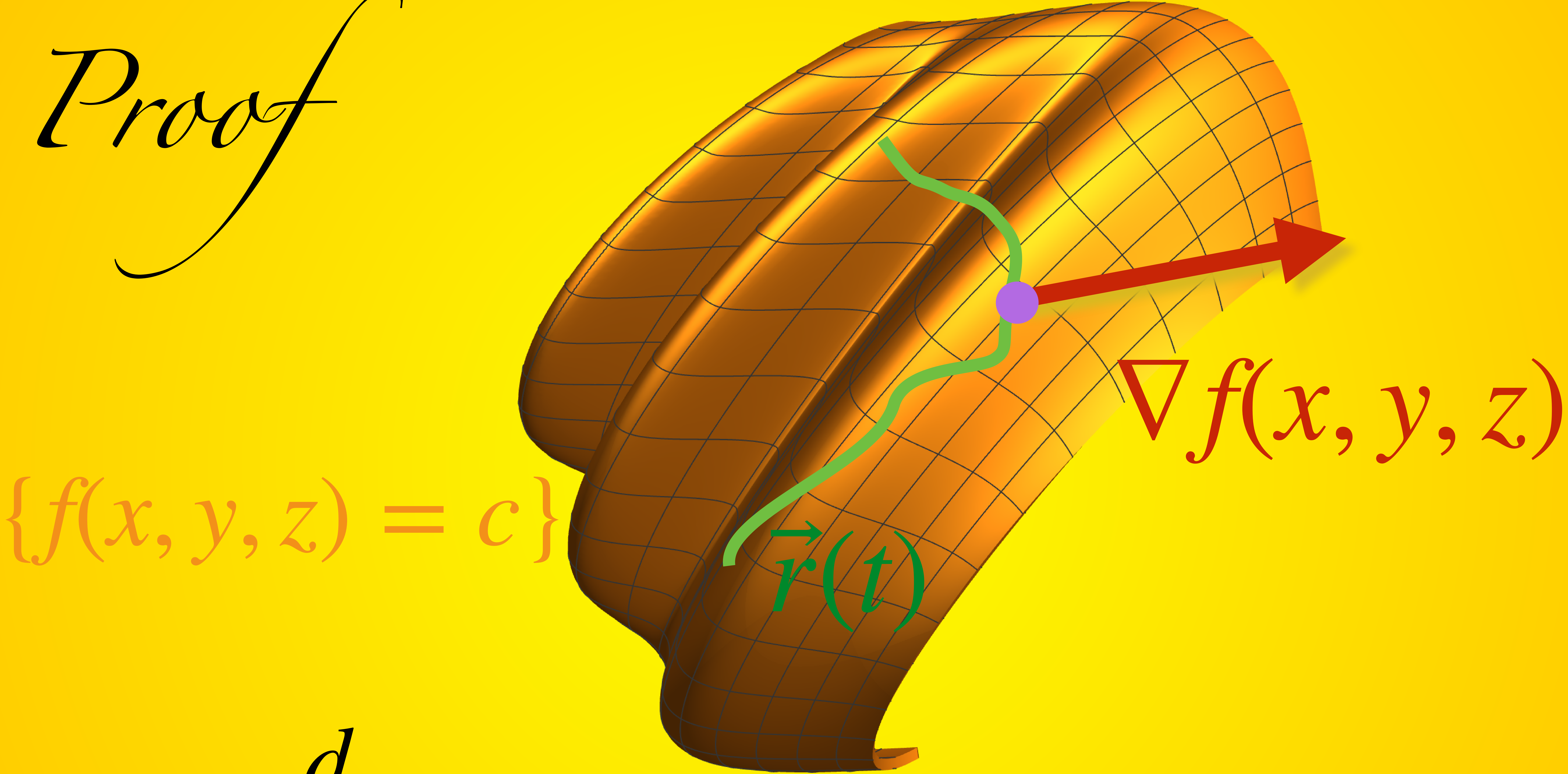
5) Tangent planes

Gradient Theorem

Gradient Theorem

∇f is perpendicular to the level surface $\{f = c\}$

Proof



$$0 = \frac{d}{dt} f(\vec{r}(t)) = \nabla f(\vec{r}(t)) \cdot \vec{r}\,'(t)$$

Steapest ascent

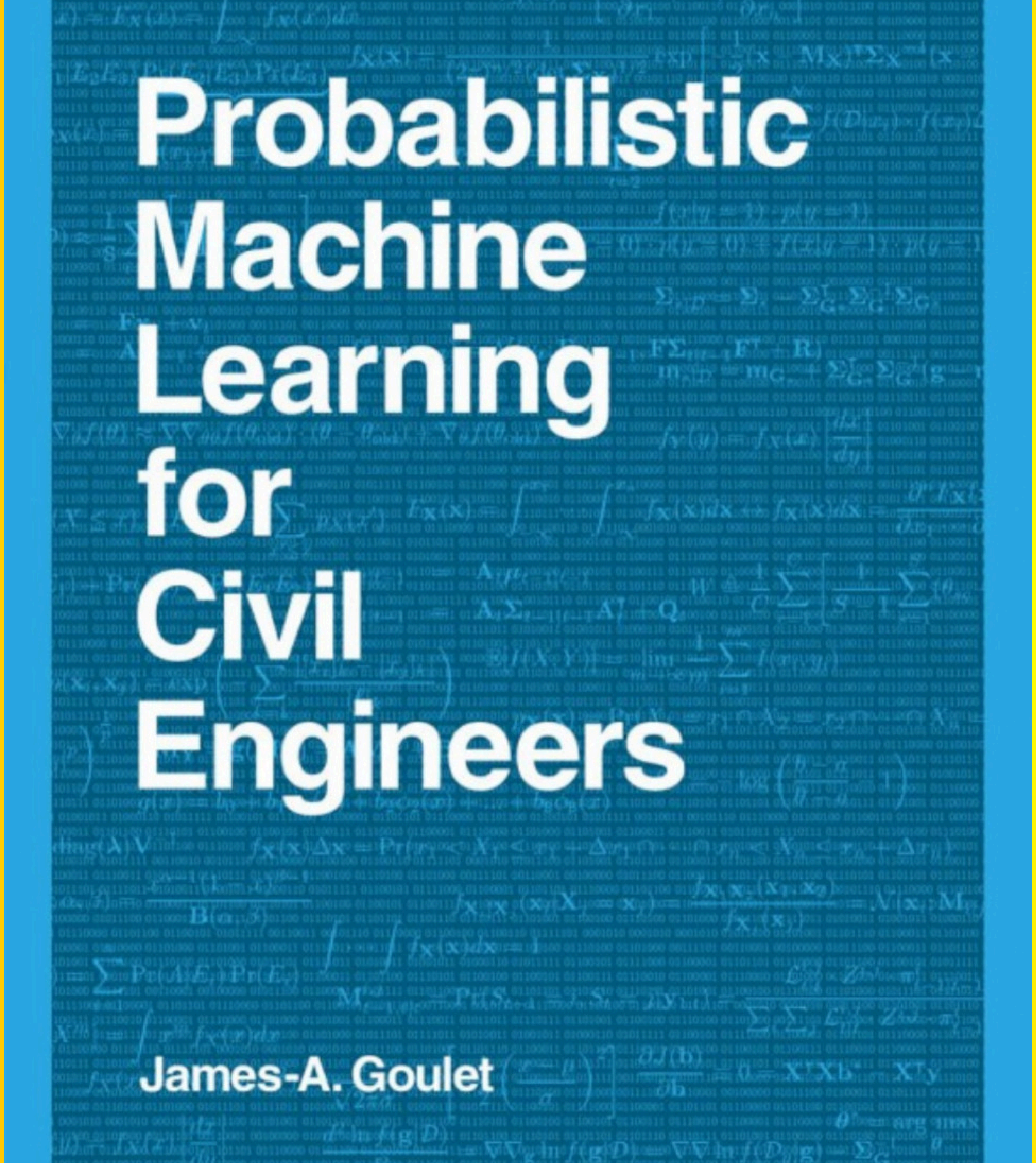
Steepest Ascent

∇f is the direction in which f increases most

Proof

$$\frac{d}{dt}f(\vec{r}(t)) = \nabla f(\vec{r}(t)) \cdot \vec{r}'(t) = |\nabla f(\vec{r}(t))| \cdot |\vec{r}'(t)| \cos(\theta)$$

is maximal for $\cos(\theta) = 1$



Algorithm 1: Gradient ascent with backtracking line search

1 initialize $\lambda = \lambda_0$, $\theta_{\text{old}} = \theta_0$, define ϵ, c, $\tilde{f}(\theta)$
2 **while** $|\tilde{f}'(\theta_{old})| > \epsilon$ **do**
3 compute $\begin{cases} \tilde{f}(\theta_{\text{old}}) & \text{(Function value)} \\ \tilde{f}'(\theta_{\text{old}}) & \text{(1}^{\text{st}}\text{ derivative)} \end{cases}$
4 compute $\theta_{\text{new}} = \theta_{\text{old}} + \underbrace{\lambda \tilde{f}'(\theta_{\text{old}})}_{d}$
5 **if** $\tilde{f}(\theta_{new}) < \tilde{f}(\theta_{old}) + c \cdot d\tilde{f}'(\theta_{old})$ **then**
6 assign $\lambda = \lambda/2$ (Backtracking)
7 Goto 4
8 assign $\lambda = \lambda_0$, $\theta_{\text{old}} = \theta_{\text{new}}$
9 $\theta^* = \theta_{\text{old}}$

Figure 5.4 presents the first two steps of the application of algorithm 1 to a non-convex/non-concave function with an initial value $\theta_0 = 3.5$ and a scaling factor $\lambda_0 = 3$. For the second step, the scaling factor λ has to be reduced twice in order to satisfy the Armijo rule. One of the difficulties with gradient ascent is that the convergence speed depends on the choice of λ_0. If λ_0 is too small, several steps will be wasted and convergence will be slow. If λ_0 is too large, the algorithm may not converge.

Figure 5.5 presents a limitation common to all convex optimization methods when applied to functions involving local maxima; if the starting location θ_0 is not located on the slope segment leading to the global maximum, the algorithm will most likely miss it and converge to a local maximum. The task of selecting a proper value $\boldsymbol{\theta}_0$ is nontrivial because in most cases, it is not possible to visualize $\tilde{f}(\boldsymbol{\theta})$. This issue can be tackled by attempting multiple starting locations $\boldsymbol{\theta}_0$ and by using domain knowledge to identify proper starting locations.

Gradient ascent can be applied to search for the maximum of a multivariate function by replacing the univariate derivative by the gradient so that

$$\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}_{\text{old}} + \lambda \cdot \nabla_{\boldsymbol{\theta}} \tilde{f}(\boldsymbol{\theta}_{\text{old}}).$$

As illustrated in figure 5.6, because gradient ascent follows the direction where the gradient is maximal, it often displays an oscillatory pattern. This issue can be mitigated by introducing a *momentum* term in the calculation of $\boldsymbol{\theta}_{\text{new}}$,[3]

$$\begin{aligned} \mathbf{v}_{\text{new}} &= \gamma \cdot \mathbf{v}_{\text{old}} + \lambda \cdot \nabla_{\boldsymbol{\theta}} \tilde{f}(\boldsymbol{\theta}_{\text{old}}), \\ \boldsymbol{\theta}_{\text{new}} &= \boldsymbol{\theta}_{\text{old}} + \mathbf{v}_{\text{new}} \end{aligned}$$

where $\mathbf{v}$ can be interpreted as a velocity that carries the momentum from the previous iterations.

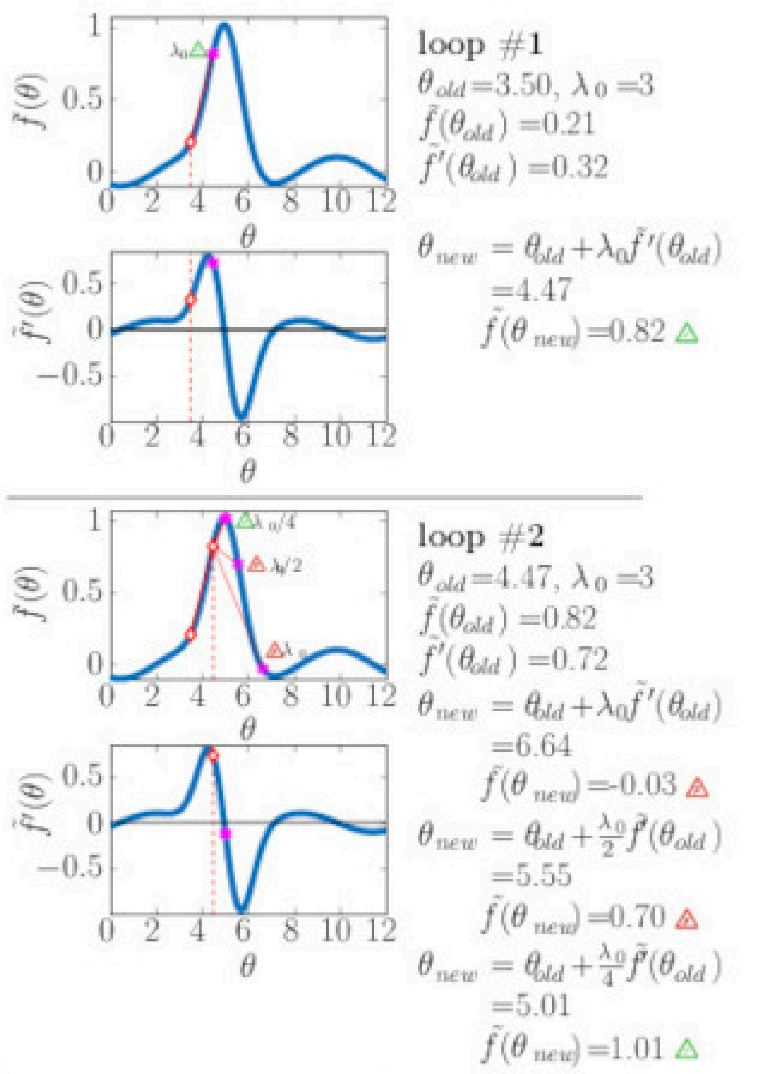


Figure 5.4: Example of application of gradient ascent with backtracking for finding the maximum of a function.

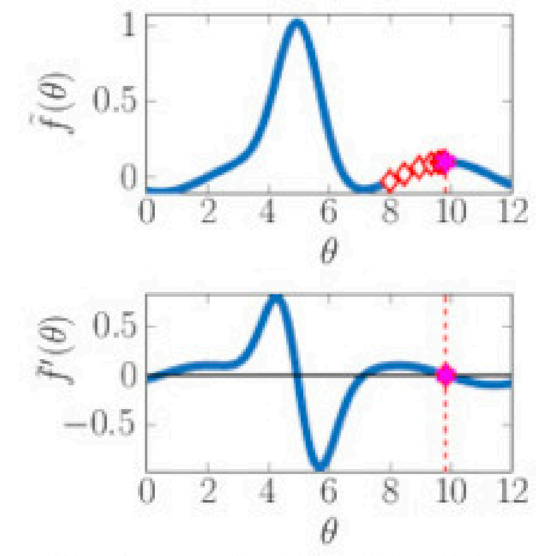


Figure 5.5: Example of application of gradient ascent converging to a local maximum for a function.

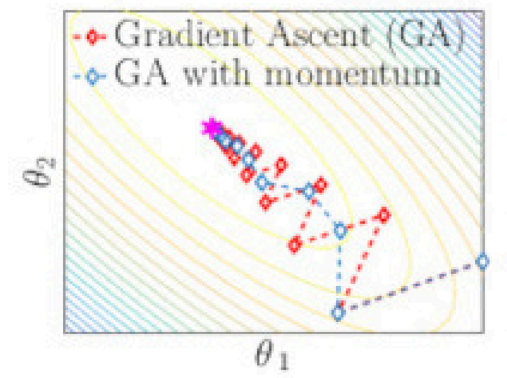


Figure 5.6: Comparison of gradient ascent with and without momentum.

[3] Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning representations by back-propagating errors. *Nature 323*, 533–536

5.1 Gradient Ascent

A *gradient* is a vector containing the partial derivatives of a function with respect to its variables. For a continuous function, the maximum is located at the point where its gradient equals zero. *Gradient ascent* is based on the principle that as long as we move in the direction of the gradient, we are moving toward a maximum. For the unidimensional case, we choose to move to a new position θ_{new} defined as the old value θ_{old} plus a search direction d defined by a scaling factor λ times the derivative estimated at θ_{old},

$$\theta_{\text{new}} = \theta_{\text{old}} + \underbrace{\lambda \cdot \tilde{f}'(\theta_{\text{old}})}_{d}.$$

A common practice for setting λ is to employ *backtracking line search* where a new position is accepted if the *Armijo rule*[2] is satisfied so that

$$\tilde{f}(\theta_{\text{new}}) \geq \tilde{f}(\theta_{\text{old}}) + c \cdot d\tilde{f}'(\theta_{\text{old}}), \text{ with } c \in (0,1). \qquad (5.2)$$

Figure 5.3 presents a comparison of the application of equation 5.2 with the two extreme cases, $c = 0$ and $c = 1$. For $c = 1$, θ_{new} is

Derivative

$$\tilde{f}'(\theta) \equiv \frac{d\tilde{f}(\theta)}{d\theta}$$

Gradient

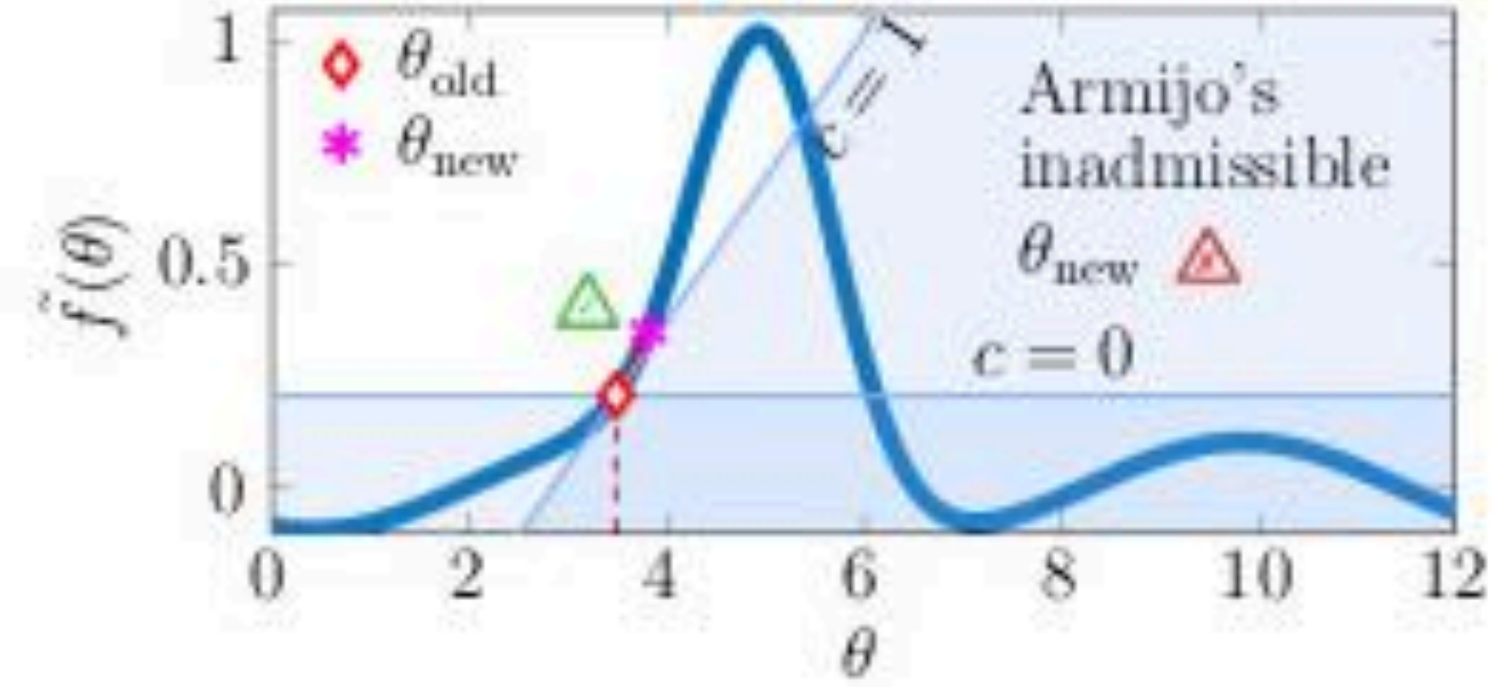
$$\begin{aligned}\nabla \tilde{f}(\boldsymbol{\theta}) &\equiv \nabla_{\boldsymbol{\theta}} \tilde{f}(\boldsymbol{\theta}) \\ &= \left[\frac{\partial \tilde{f}(\boldsymbol{\theta})}{\partial \theta_1} \; \frac{\partial \tilde{f}(\boldsymbol{\theta})}{\partial \theta_2} \; \cdots \; \frac{\partial \tilde{f}(\boldsymbol{\theta})}{\partial \theta_n}\right]^\intercal\end{aligned}$$

Maximum of a concave function

$$\theta^* = \arg\max_{\theta} \tilde{f}(\theta) : \frac{d\tilde{f}(\theta^*)}{d\theta} = 0$$

λ is also known as the *learning rate* or *step length*.

[2] Armijo, L. (1966). Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics 16*(1), 1–3



Mt. Ephraim
Reservoirs
Turkey Hill
300
200
100
ARLINGTON
ARLINGTON
Arlington Heights
Arlington
BOSTON & MAINE
MYSTIC
Brook
Pond

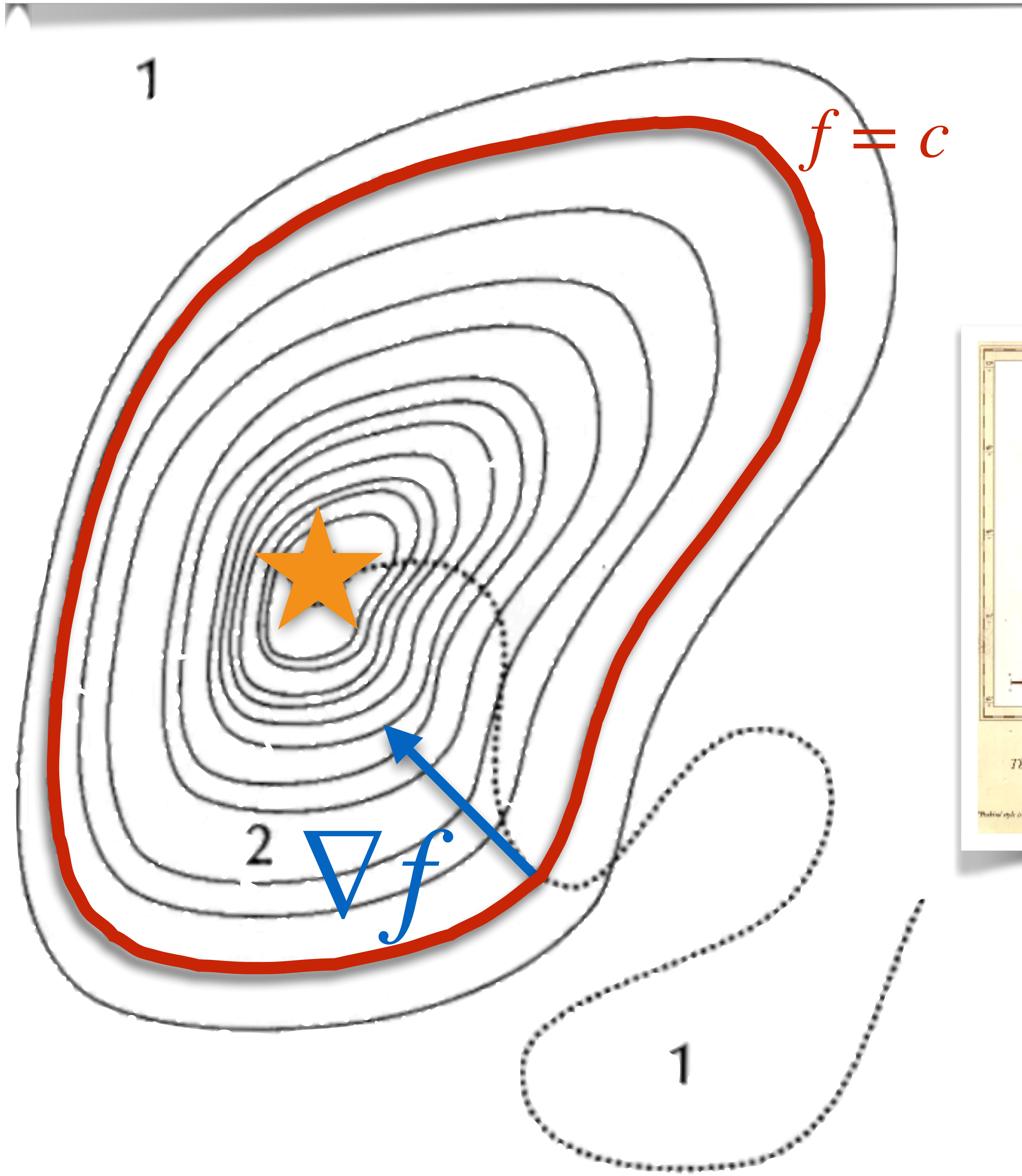

Turkey
Hill
ARLINGTON
RLINGTON
Arlington
Heights
100
200


Grindelwald
Eiger
Wärgistal
Alpiglen
Challifirn

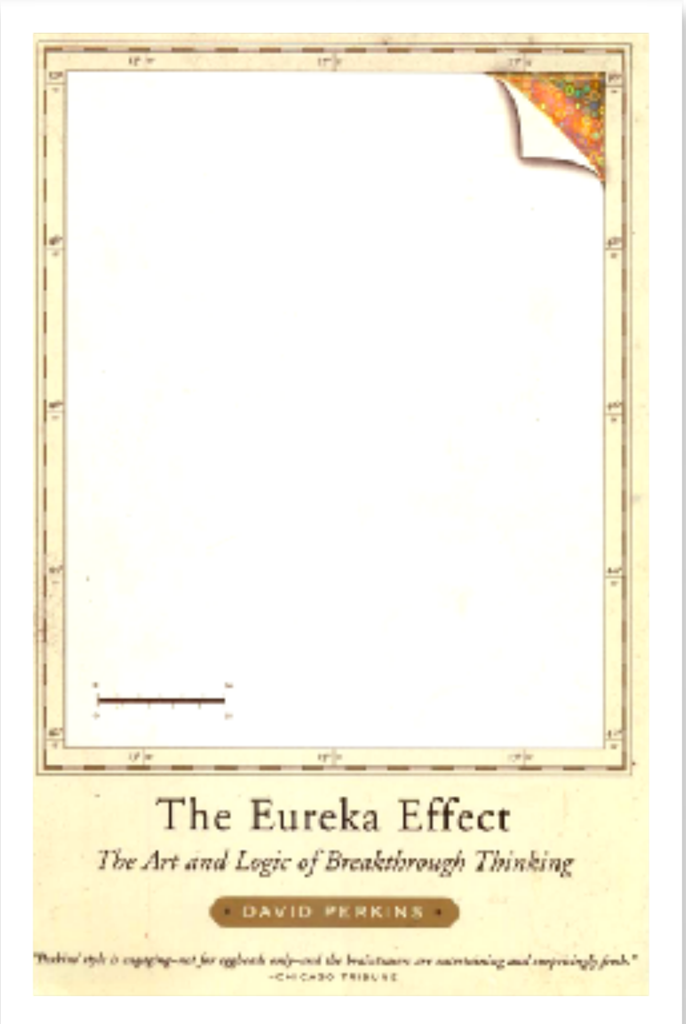
1846
Im glatten Wang
2287
Hönysch
2134
Chräjenbiel
2802
3355
Mittellegi
2200
Eigerwand
2561
3692
2865
Wart
3066
2663
3970
Eiger
Rotstock
Obers Challiba
Eigergletscher
Kl. Eiger
3614
Eismeer
3159.7
2600
Eigerjoch
3472
3400
2800
Eigergletscher
Berglihütte
3299

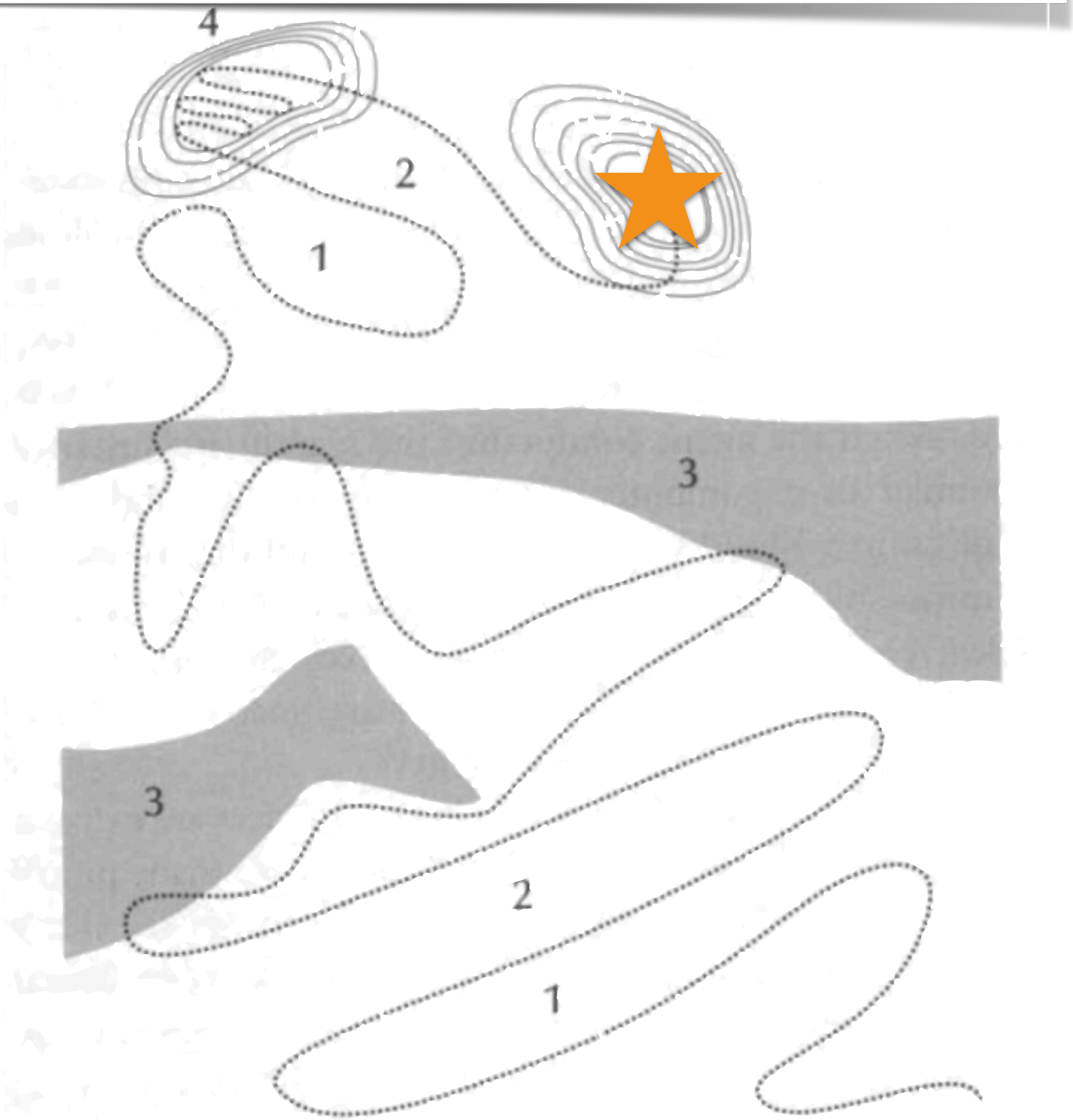

Alcohol
& calculus
don't mix.
Never drink
& derive.
$r(t) = x + t\nabla f(x)$
$\frac{d}{dt} f(x) = \nabla f(x) \cdot \nabla f(x)$

From D. Perkins: The Eureka Effect, the Art and Logic of Breakthrough Thinking , 2000



Search in a Homing Space: 1. Clueless regions.
2. Large clued regions leading to the target.





Search in a Klondike Space: 1. A large space wtih few solutions (a wilderness trap). 2. Regions with no clues pointing direction (plateau traps). 3. A barrier isolates the solution (creating a canyon trap). 4. An area of high promise but no solution (an oasis trap).

Quadratic Approximation

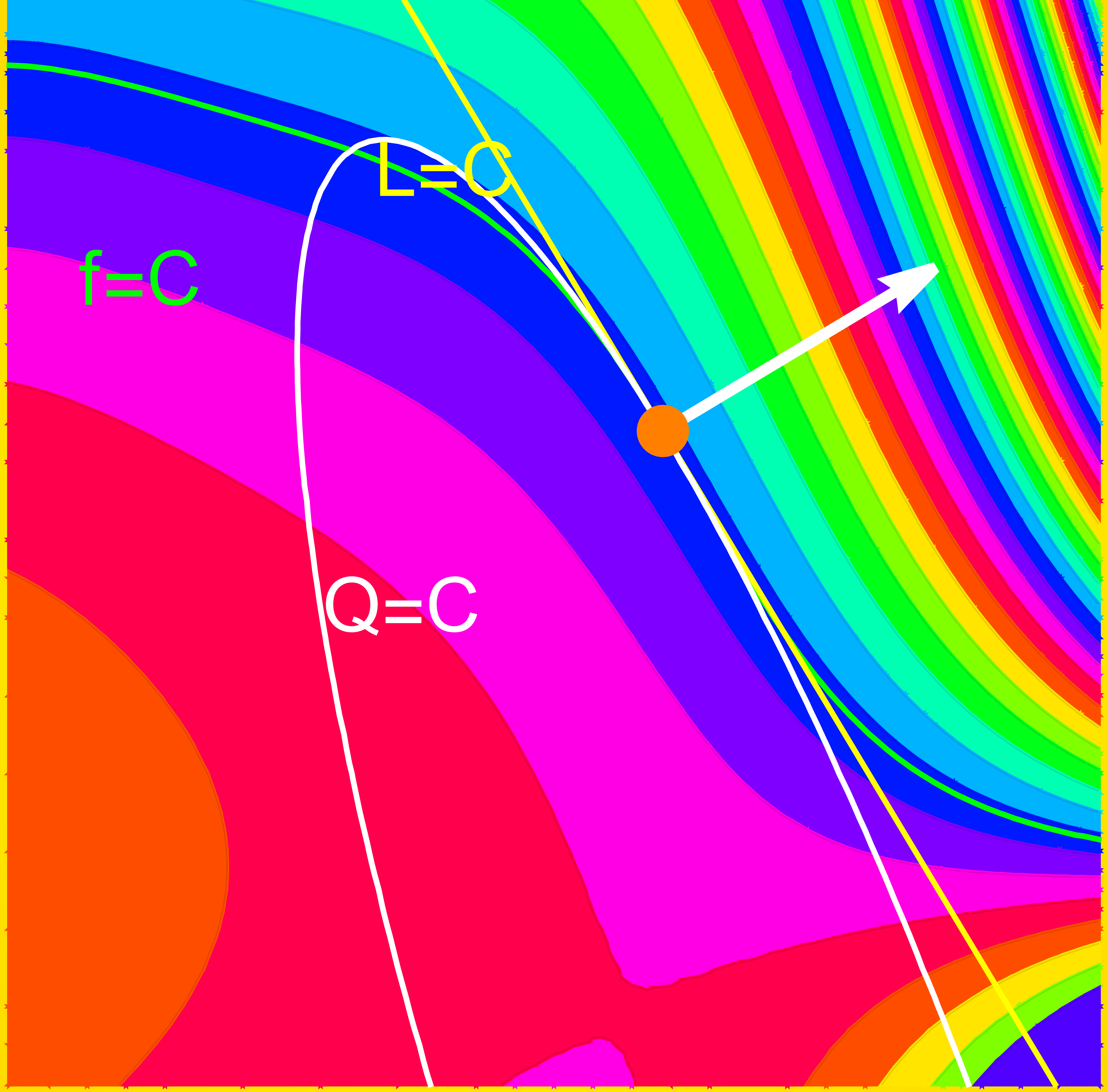
Quadratic Approximation

$$Q(x,y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$$

$$\frac{f_{xx}(x_0, y_0)(x - x_0)^2 + f_{yy}(x_0, y_0)(y - y_0^2) + 2f_{xy}(x_0, y_0)(x - x_0)(y - y_0)}{2}$$

The best quadratic function near the point (x_0, y_0)

2D



$f(x, y) = x^3 y$
$\nabla f(1,2) = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$
$(x_0, y_0) = (1,2)$
$r(t) = x + t\nabla f(x)$
$\frac{d}{dt} f(x) = \nabla f(x) \cdot \nabla f(x)$
$L(x, y) = 2 + 6(x - 1) + (y - 2)$


3D
f=C
Q=C
L=C

THE END